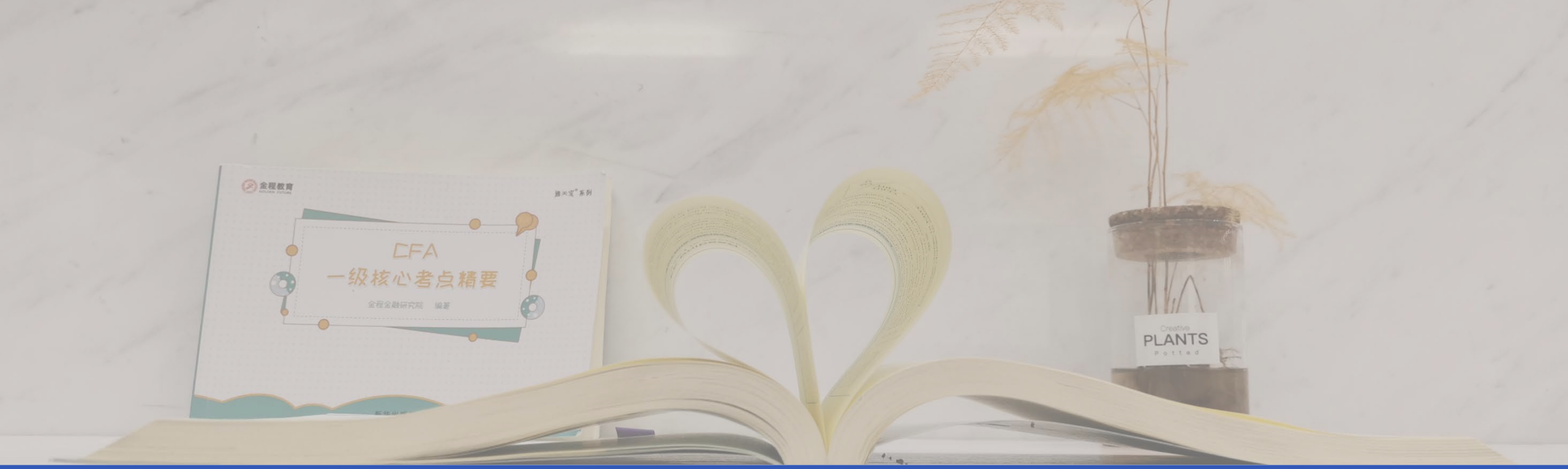


Quantitative Methods

CFA一级培训项目

讲师：王慧琳

师资介绍

1. 基本介绍

金程教育资深培训师、上海财经大学经济学学士、美国约翰霍普金斯大学金融学硕士、CFA、FRM、ESG investing持证人

2. 工作背景

多家知名机构内训项目授课，参与出版CFA相关系列丛书教材。本科毕业于上海财经大学，研究生毕业于约翰霍普金斯大学，一次性通过CFA一二三级考试，对于考试重点和应试技巧有自己的心得。

3. 服务客户

中国工商银行、中国银行、建设银行、农业银行、杭州银行、兴业证券、南京证券、湘财证券、兴业银行、中国人寿、人保资产管理、中国平安、民生银行、华夏基金、中邮基金、富国基金、中国再保险、中国进出口银行等。

Topic Weightings in CFA Level I

Topics	Weights (%)
Quantitative Methods	8-12
Economics	8-12
Financial Statement Analysis	13-17
Corporate Issuers	8-12
Equity	10-12
Fixed Income	10-12
Derivatives	5-8
Alternative Investments	5-8
Portfolio Management	5-8
Ethical and Professional Standards	15-20

Quantitative Methods

1. Rates and Returns
2. The Time Value of Money in Finance
3. Statistical Measures of Asset Returns
4. Probability Trees and Conditional Expectations
5. Portfolio Mathematics
6. Simulation Methods
7. Estimation and Inference
8. Hypothesis Testing
9. Parametric and Non-Parametric Tests of Independence
10. Simple Linear Regression
11. Introduction to Big Data Techniques

中文精读

1. 利率和收益
2. 金融中的货币时间价值
3. 资产收益率的统计度量
4. 概率树和条件期望
5. 概率论基础
6. 模拟方法
7. 抽样和估计
8. 假设检验
9. 独立性的参数检验与非参数检验
10. 线性回归分析
11. 大数据分析

Framework

Module

Rates and Returns

1. Interest rates and time value of money
2. Annualized return
3. Average returns
4. Money-weighted and time-weighted return
5. Other major return measures and their Applications

Interest rates and time value of money

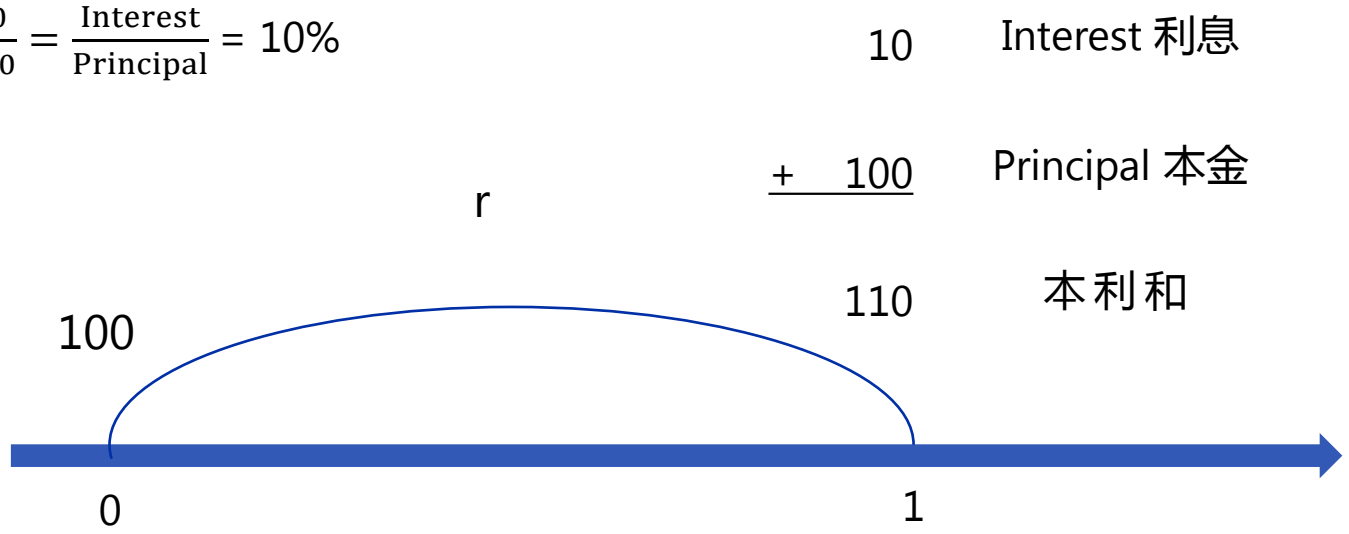
- Interpret interest rates as required rates of return, discount rates, or opportunity costs and explain an interest rate as the sum of a real risk-free rate and premiums that compensate investors for bearing distinct types of risk

What's an interest rate?

- **Interest rate**
  - Interest rate (r) measures the amount of money (interest) per unit of investment (principal).
  - A rate of return that reflects the relationship between differently dated cash flows.

$$100 + 100 \times 10\% = 100 \times (1 + 10\%) = 110$$

$$r = \frac{10}{100} = \frac{\text{Interest}}{\text{Principal}} = 10\%$$



Interest rate

- **Required rate of return is**
  - affected by the **supply and demand** of funds in the market;
  - the minimum rate of return an investor must receive to accept the investment.
  - usually for particular investment.
- **Discount rate is**
  - the interest rate we use to **discount** payments to be made in the future.
  - usually used interchangeably with the interest rate.
- **Opportunity cost is**
  - also understood as a form of interest rate. It is the value that investors **forgo** by choosing a particular course of action.

Determinants of Interest Rates

- **Interest rate**

Required interest rate on a security (R_i)

= *nominal risk-free rate (Rf)* → *nominal risk-free rate = real risk-free rate + expected inflation rate*

\+ risk premiums (RP)

Tips: Nominal rate = real rate + expected inflation rate

\+ default risk premium (DRP) → the risk of not making the promised payments.

\+ liquidity risk premium (LRP) → the risk of not converting to cash in fair value.

\+ maturity risk premium (MRP) → the risk of increasing volatility with extended period.

Example

Determinants of Interest Rates

- The sum of the real risk-free interest rate and the expected inflation premium is best described as:

A. the nominal risk-free interest rate.

B. the default risk premium.

C. the liquidity premium.

- **Correct Answer: A.**

Example

- The table below gives current information on the interest rates for two two-year and two eight-year maturity investments. The table also gives the maturity, liquidity, and default risk characteristics of a new investment possibility (Investment 3). All investments promise only a single payment (a payment at maturity). Assume that premiums relating to inflation, liquidity, and default risk are constant across all time horizons.

Investment	Maturity(in years)	Liquidity	Default Risk	Interest rate(%)
1	2	High	Low	2.0
2	2	Low	Low	2.5
3	7	Low	Low	r_3
4	8	High	Low	4.0
5	8	Low	High	6.5

- Based on the information in the above table, address the following questions A, B and C:

A. Explain the difference between the interest rates on Investment 1 and Investment 2.
B. Estimate the default risk premium.
C. Calculate upper and lower limits for the interest rate on Investment 3, r_3.

Example

- **Correct Answer:**
  - **QA**
    - Investment 2 is identical to Investment 1 except that Investment 2 has low liquidity.
    - liquidity premium=r_2-r_1, which represents compensation for the risk of loss relative to an investment's fair value if the investment needs to be converted to cash quickly.
  - **QB**
    - To estimate the default risk premium, find the two investments that have the same maturity but different levels of default risk. Both Investments 4 and 5 have a maturity of eight years. Investment 5, however, has low liquidity and thus bears a liquidity premium. The difference between the interest rates of Investments 5 and 4 is 2.5 percentage points. The liquidity premium is 0.5 percentage point (from Part A). This leaves 2.5 − 0.5 = 2.0 percentage points that must represent a default risk premium reflecting Investment 5's high default risk.

Example

- **Correct Answer:**
  - **QC**
    - Investment 3 has liquidity risk and default risk comparable to Investment 2, but with its longer time to maturity, Investment 3 should have a higher maturity premium. The interest rate on Investment 3, r_3, should thus be above 2.5 percent (the interest rate on Investment 2).
    - If the liquidity of Investment 3 were high, Investment 3 would match Investment 4 except for Investment 3's shorter maturity. We would then conclude that Investment 3's interest rate should be less than the interest rate on Investment 4, which is 4 percent.
    - In contrast to Investment 4, however, Investment 3 has low liquidity. It is possible that the interest rate on Investment 3 exceeds that of Investment 4 despite 3's shorter maturity, depending on the relative size of the liquidity and maturity premiums. However, we expect r_3 to be less than 4.5 percent, the expected interest rate on Investment 4 if it had low liquidity.
    - Thus 2.5 percent $< r_3 <$ 4.5 percent.

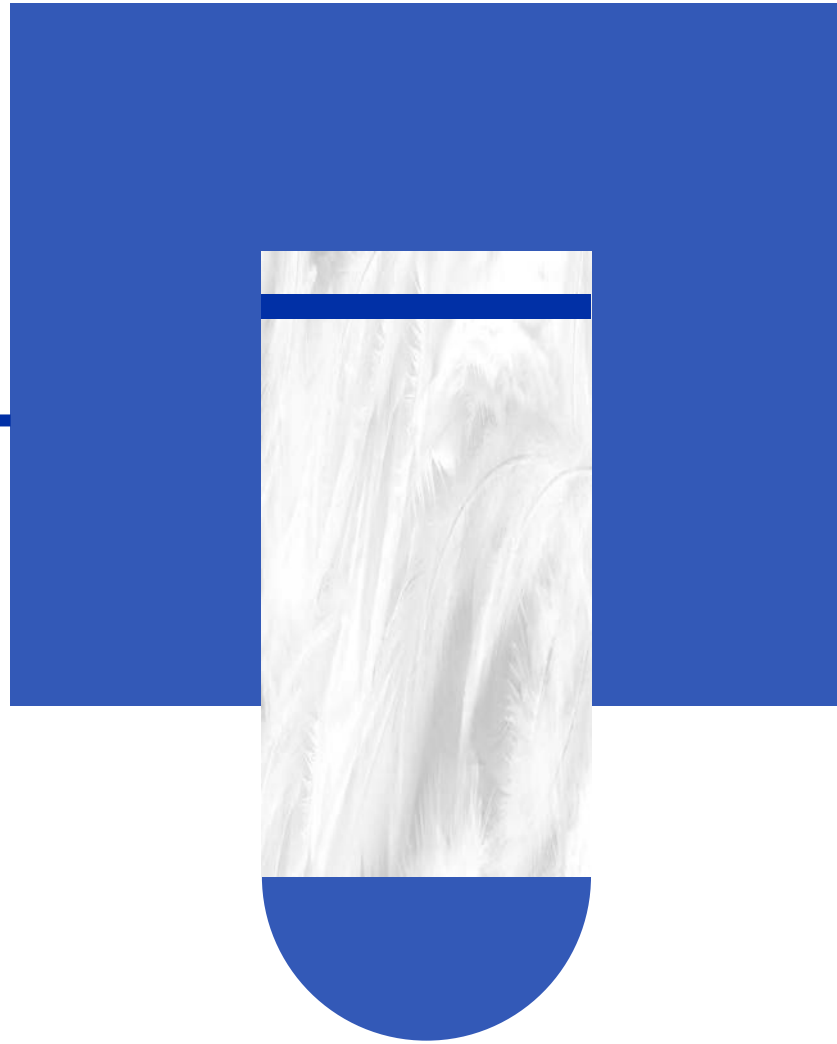
Summary

Rates and Returns

Interest rates and time value of money

Annualized return

- Calculate and interpret annualized return measures and continuously compounded returns, and describe their appropriate uses



Holding Period Return

- **Holding period return (HPR)**
  - HPR is simply the percentage change in the value of an investment over the period it is hold.

$$\text{HPR} = \frac{P_1 - P_0 + CF_1}{P_0} \qquad \text{HPR} = \frac{\text{FV-PV}}{\text{PV}}$$

Example

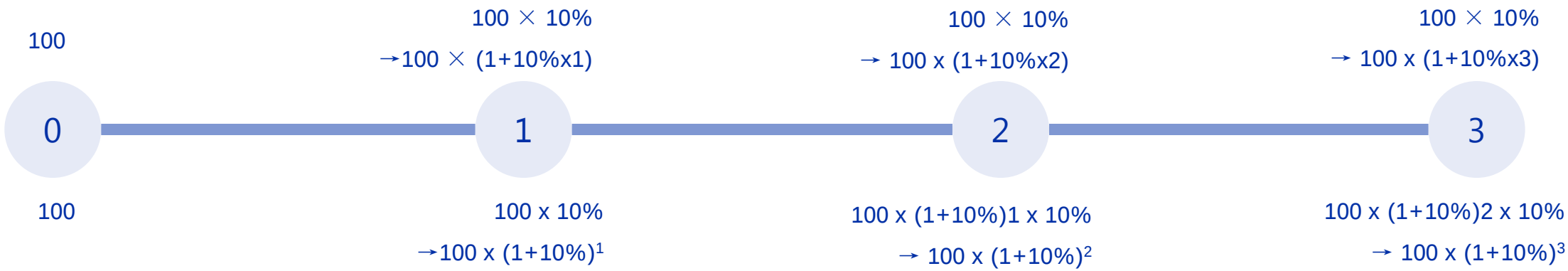
Jane Peebles purchased a T-bill that matures in 200 days for $975. The face value of the bill is $1,000. What's the holding period return of the bond?

Correct Answer:

$$HPR = \frac{FV - PV}{PV} = \frac{1000 - 975}{975} = 2.564\%$$

Compounding pattern

- **Simple Interest (on the basis of a 360-day year)**
  - There is no opportunity to re-invest the interest payments during the life of an investment and thereby earn extra income.



- **Compound interest (on the basis of a 365-day year)**
  - Interest amounts will be received periodically and can be re-invested (usually at the same rate).

Annualizing rate

- **Effective Annual Rate (EAR) calculation**

$$\text{EAR}=(1+\text{periodic rate})^{m}-1 \quad \longleftrightarrow \quad 1+EAR=\left(1+\frac{r}{m}\right)^{m}$$

  - If semi-annually compounding, then m=2
  - If quarterly compounding, then m=4
  - If continuously compounding, then EAR = $e^{\text{annual int}}-1$

- **Tips**
  - Calculation: calculate EAR, or calculate the frequency of compounding
  - Feature
    - The more frequency of compounding, the larger the EAR.
    - The largest EAR exists if it is continuously compounding.

Example

Effective Annual Rate

- A money manager has $1,000,000 to invest for one year. She has identified two alternative one-year certificates of deposit (CD) shown below:

	Compounding frequency	Annual interest rate
CD1	Quarterly	4.00%
CD2	Continuously	4.95%

Which CD has the higher effective annual rate (EAR) and how much interest will it earn?

	Higher EAR	Interest earned
A.	CD1	$41,902
B.	CD1	$40,604
C.	CD2	$50,746

- **Correct Answer: C.**

Using EAR as discount rate

- **Future value (FV):** Amount to which investment grows after one or more compounding periods.
- **Present value (PV):** Current value of some future cash flow.
  - If interests are compounded m times per year, and invest 1 year:

    $FV=PV(1+r/m)^m$
  - If interests are compounded m times per year, and invest n years:

    $FV=PV(1+ r/m)^{mn}$

    Where: m is the compounding frequency;

    r is the nominal/quoted annual interest rate.
- When we calculate the future value of continuously compounding, the formula is:

$$FV=PV\lim_{m\to\infty}(1+\frac{r}{m})^{nm}=PVe^{n\,r}$$

Example

Using EAR as discount rate

- A stated annual rate is 10%, compounded quarterly. Calculate the FV of a $200 investment at the end of two years.

- **Correct Answer:**
  - r=10%, m=4, n=1, EAR=$(1+ 10\%/4)^4-1=10.3813\%$
  - Enter relevant data for calculate.
    - N=2; I/Y=10.3813; PV=–200; PMT=0; CPT→FV=$243.68
  - Note the negative sign on PV. This is not necessary, but it makes the FV come out as a positive number. If you enter PV as a positive number, ignore the negative sign that appears on the FV.
  - This relatively simple problem could also be solved using the following equation: $FV = 200\times (1 + 10\%/4)^{4\times 2} = \$243.68$; enter 1.025 [$y^x$] 8 [×] 200 [=].

Non-annual Compounding

- In general, the formula for present value with more than one compounding period in a year:

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

  - where: m = number of compounding periods per year; R_s = quoted annual interest rate; N = number of years.

Example

A fund must make a lump-sum payment of CAD5 million 10 years from today. If the current interest rate is 6 percent a year, compounded monthly, how much should the fund invest today?

Correct Answer:

$$PV=5{,}000{,}000\times(1+\frac{6\%}{12})^{-12\times10}=2{,}748{,}164$$

Continuously Compounded Rates of Return

- The **continuously compounded return** associated with a holding period return is the natural logarithm of 1 plus that holding period return, or equivalently, the natural logarithm of the ending price over the beginning price (the **price relative**).
  - For a stock with a price relative, $\frac{S_{t+1}}{S_t} = 1 + \mathrm{HPR}_{t,t+1} = e^{r_{t,t+1}}$;
    - $r_{t,t+1} = \mathrm{Ln}(1 + \mathrm{HPR}_{t,t+1}) = \mathrm{Ln}\left(\frac{S_{t+1}}{S_t}\right)$
  - A key assumption in many investment applications is that returns are **independently and identically distributed (i.i.d.).**

$$\frac{S_T}{S_0} = \frac{S_T}{S_{T-1}} \times \frac{S_{T-1}}{S_{T-2}} \times \cdots \times \frac{S_1}{S_0} \longrightarrow 1 + HPR_{0,T} = (1 + HPR_{T-1,T-2}) \times (1 + HPR_{T-2,T-1}) \times \cdots \times (1 + HPR_{0,1})$$

$$\frac{S_T}{S_0} = \frac{S_T}{S_{T-1}} \times \frac{S_{T-1}}{S_{T-2}} \times \cdots \times \frac{S_1}{S_0} \longrightarrow r_{0,T} = r_{T-1,T-2} + r_{T-2,T-1} + \cdots + r_{0,1}$$

Example

Compare the relative performance

- An analyst seeks to evaluate three securities in portfolio for different periods of time.

Securities' Performance data		
Security	Holding period	Holding period return (HPR)
A	149 days	4.91
B	9 weeks	1.90
C	19 months	19.35

- Correct Answer: A.
  - To facilitate comparison, the three securities' returns need to be annualized:
    - Stock A annualized return = $(1.0491^{365/149}) - 1 = 12.46\%$ (Security A generated the highest annualized return.)
    - Stock B annualized return = $(1.0190^{52/9}) - 1 = 11.49\%$
    - Stock C annualized return = $(1.1935^{12/19}) - 1 = 11.82\%$

Summary

Rates and Returns

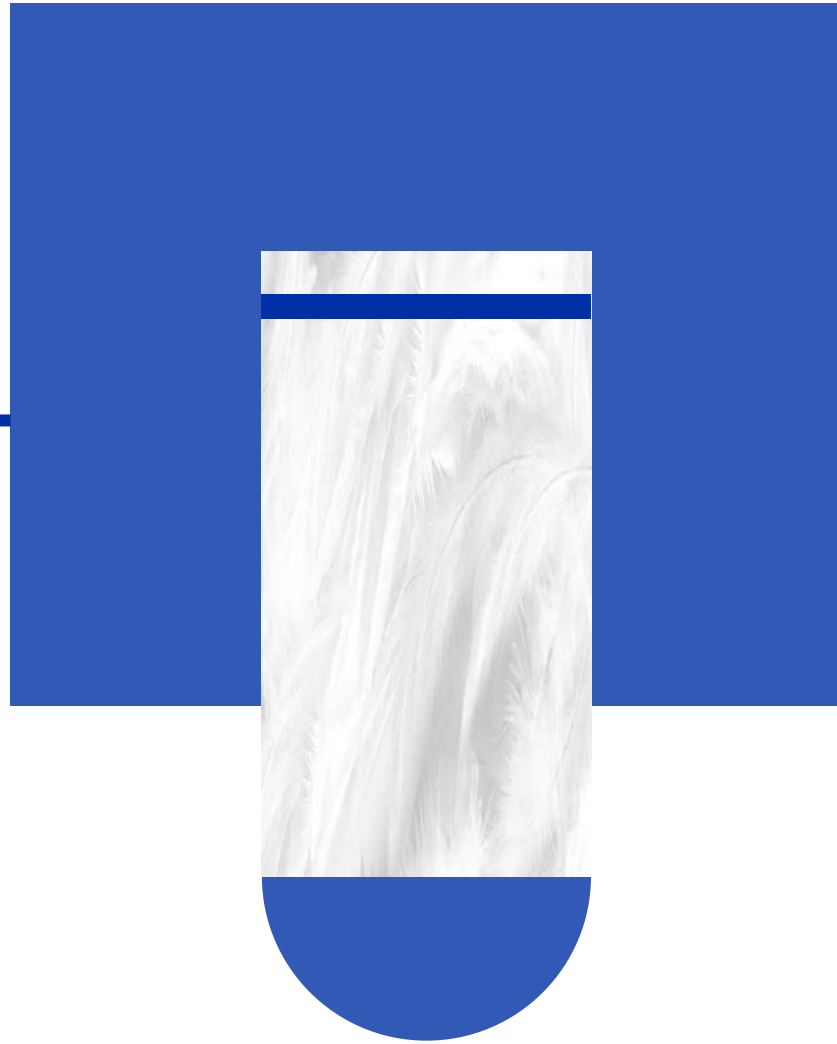
Holding period rate

Effective annual rate

Non-annual compounding

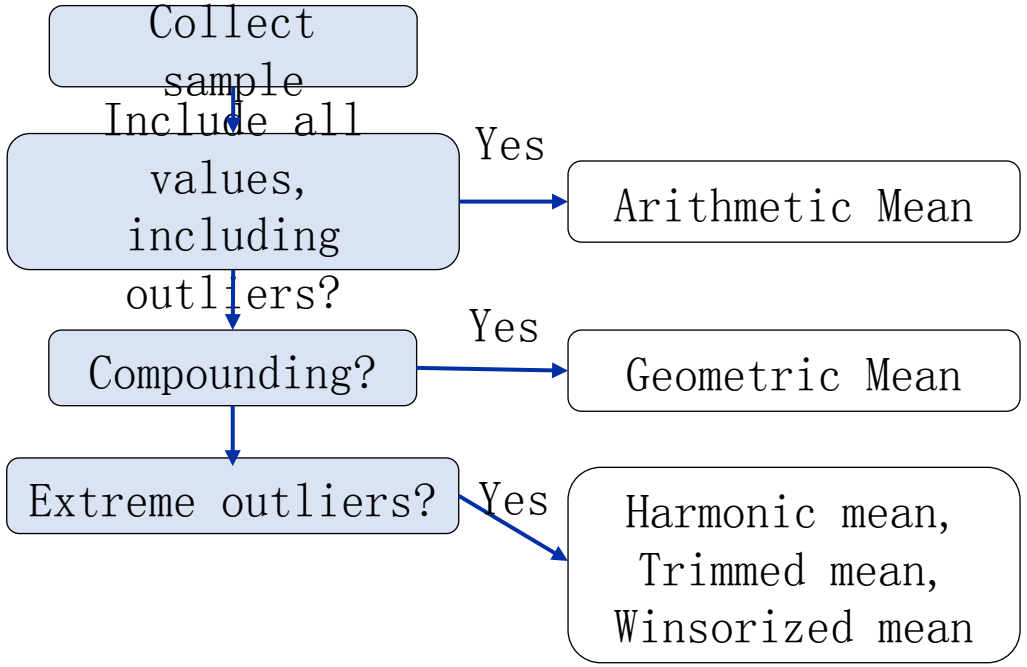
Average returns

- Calculate and interpret different approaches to return measurement over time and describe their appropriate uses



Means

- **Arithmetic Mean:** $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
- **Geometric Mean:** $G = \sqrt[N]{X_1 X_2 X_3 \ldots X_N} = (\prod_{i=1}^{N} X_i)^{1/N}$
- **Harmonic Mean:** $\bar{X}_H = \frac{n}{\sum_{i=1}^{n} (1/X_i)}$
- Harmonic Mean≤ Geometric Mean ≤ Arithmetic Mean
- Arithmetic mean × Harmonic mean ≈ Geometric mean^2



- **Trimmed mean**: remove a small defined percentage of the largest and smallest values and calculate the mean by averaging the remaining observations
- **Winsorized mean**: replace extreme values at both ends with the values of their nearest observations and calculate the mean by averaging the remaining observations.

Example

Geometric and Arithmetic Mean Returns

- Calculate the arithmetic and geometric mean returns over the three years for the following three stock indexes: Country D, Country E, and Country F.

	Annual Return (%)			Sum	Arithmetic Mean Returns
	Year 1	Year 2	Year 3		
country D	-2.40%	-3.10%	6.20%	0.70%	0.233%
country E	-4.00%	-3.00%	3.00%	-4.00%	-1.333%
country F	5.40%	5.20%	-1.00%	9.60%	3.200%

	1+return in decimal form (1+Rt)			Product	3rd root	Geometric Mean Returns
	Year 1	Year 2	Year 3			
country D	0.976	0.969	1.062	1.00438	1.001	0.146%
country E	0.960	0.970	1.030	0.95914	0.986	-1.381%
country F	1.054	1.052	0.990	1.09772	1.032	3.157%

Example

The Harmonic Mean

- Calculate the harmonic mean returns over the three years for the following three stock indexes: Country D, Country E, and Country F.

	1+return in decimal form (1+Rt)			Reciprocal			Sum	Reciprocal	product n	Return
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3				
country D	0.976	0.969	1.062	1.025	1.032	0.942	2.998	0.334	1.001	0.06%
country E	0.960	0.970	1.030	1.042	1.031	0.971	3.043	0.329	0.986	-1.43%
country F	1.054	1.052	0.990	0.949	0.951	1.010	2.909	0.344	1.031	3.11%

- Harmonic Mean≤ Geometric Mean ≤ Arithmetic Mean
  - For country D: 0.06%<0.146%<0.233%
- Arithmetic mean × Harmonic mean ≈ Geometric mean^2
  - For country D: (1+0.233%) x (1+0.06%) ≈ (1+0.146%)^2

Example
Calculating the Arithmetic, Geometric, and Harmonic Means

- Each year in Dec, a securities analyst selects 10 favorite stocks for the next year. The following exhibit gives the P/Es, the ratio of share price to projected earnings per share (EPS), for top-10 stock picks for the next year. For these 10 stocks, calculate the arithmetic mean P/E, geometric mean P/E, harmonic mean P/E, a 20% trimmed mean and a 80% winsorized mean.

Stock	P/E	Natural log of the P/E: $\ln(X_i)$	Reciprocal ($1/X_i$)
Stock 1	22.29	3.104	0.045
Stock 2	15.54	2.743	0.064
Stock 3	9.38	2.239	0.107
Stock 4	15.12	2.716	0.066
Stock 5	10.72	2.372	0.093
Stock 6	14.57	2.679	0.069
Stock 7	7.20	1.974	0.139
Stock 8	7.97	2.076	0.125
Stock 9	10.34	2.336	0.097
Stock 10	8.35	2.122	0.120
Sum	121.4800	24.3613	0.9247

Example
Calculating the Arithmetic, Geometric, and Harmonic Means

- **Correct Answer:**
  - Arithmetic mean P/E is 121.48/10 = 12.1480.
  - Geometric mean P/E is 11.4287.
    - $(22.29x15.54x...x1034x8.35)^{1/10}$
    - $e^{[\ln(22.29x15.54x...x1034x8.35)]/10} = e^{24.3613/10}$
      - $X_G = (X_1X_2...X_n)^{1/n} \rightarrow \ln X_G = [\ln(X_1X_2...X_n)]/n \rightarrow X_G = e^{[\ln(X1X2...Xn)]/n}$
  - Harmonic mean P/E is 10/0.9247 = 10.8142.

Example
Calculating the trimmed mean and winsorized means

- A 20% trimmed mean discards the lowest 10% and the highest 10 % of P/E values and computes the mean of the remaining 80% of P/E values. (E.g., sports competitions when judges' lowest and highest scores are discarded in computing a contestant's score.)
- A 80% winsorized mean is calculated after assigning one specified low value (bottom 10% values) to a stated percentage of the lowest values (10th percentile) in the dataset and one specified high value (top 10% value) to a stated percentage of the high[redacted]values (90th percentile) in the dataset.

Stock	P/E	Trim[redacted]mean	Winsorized mean
Stock 7	7.20		7.97
Stock 8	7.97	7.97	7.97
Stock 10	8.35	8.35	8.35
Stock 3	9.38	9.38	9.38
Stock 9	10.34	10.34	10.34
Stock 5	10.72	10.72	10.72
Stock 6	14.57	[redacted]	14.57
Stock 4	15.12		15.12
Stock 2	15.54	[redacted]	15.54
Stock 1	22.29	22.29	15.54
Average	12.15	11.50	11.55

Trimmed mean and winsorized means are lower than arithmetic mean

Summary

Rates and Returns

Arithmetic mean

Geometric mean

Harmonic mean

Trimmed & Winsorized mean

Money-weighted and time-weighted return

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

Time-Weighted Rate of Return

- **Time-weighted Rate of Return (TWRR)**
  - Time-weighted rate of return measures the compound rate of growth.
  - Calculation
    - ✓ Firstly, calculate the HPR on the portfolio for each subperiod;
    - ✓ Then, compute the annualized TWRR.
      - $TWRR = \sqrt[n]{\prod_{i=1}^{N}(1 + HPR_i)}$ -1,
        - where n=number of years;
        - N=number of periods.

Example

Time-Weighted Rate of Return

- Assume an investor purchases a share of stock for $50 at time t = 0, and another share at $65 at time t = 1, and at the end of Year 1 and Year 2, the stock paid a $2 dividend. Also, at the end of Year 2, the investor sold both shares for $70 each. The time-weighted rate of return on the investment is:

  A. 18.27%.

  B. 20.13%.

  C. 21.83%.

- **Solution: C.**

  $HPR_1 = (65+2)/50-1=34\%$, $HPR_2 = (140+4)/130-1=10.77\%$

  Time-weighted return $= [(1+34\%) \times (1+10.77\%)]^{0.5} - 1 = 21.83\%$

Money-Weighted Rate of Return

- **Money-weighted Rate of Return (MWRR)**
  - The **IRR** based on the cash flows related to the investment.
  - Calculation
    - ✓ Firstly, determine the timing of each cash flow;
    - ✓ then, using the calculator to compute IRR, or using geometric mean.

Example

Money-Weighted Rate of Return

- Assume an investor purchases a share of stock for $50 at time t = 0, and another share at $65 at time t = 1, and at the end of Year 1 and Year 2, the stock paid a $2 dividend. Also, at the end of Year 2, the investor sold both shares for $70 each. The money-weighted rate of return on the investment is:

  A. 15.45%.

  B. 16.73%.

  C. 18.02%.

- Solution: C.

  CF_0=-50, CF_1=-65+2=-63, CF_2=(70+2)×2=144

  Calculate IRR=18.02%

TWRR vs. MWRR

- **The relationship between TWRR and MWRR**
  - Both TWRR and MWRR are **annual rates**.
  - Time-weighted return **is not influenced by cash flow**, but money-weighted return will be affected by cash flow.

NPV & IRR

- **Net present value (NPV)**
  - ✓ $\mathrm{NPV} = \mathrm{CF}_0 + \frac{\mathrm{CF}_1}{(1+r)^1} + \frac{\mathrm{CF}_2}{(1+r)^2} + \cdots + \frac{\mathrm{CF}_n}{(1+r)^n}$
    - PV of the future after-tax cash flows minus the investment outlay
    - r: required rate of return (opportunity cost of capital, COC) related with risks
- **Internal rate of return (IRR)**
  - ✓ Discount rate that makes the PV of the future after-tax cash flows equal that investment outlay (NPV=0).
    - $\mathrm{CF}_0 + \frac{\mathrm{CF}_1}{(1+\mathrm{IRR})^1} + \frac{\mathrm{CF}_2}{(1+\mathrm{IRR})^2} + \cdots + \frac{\mathrm{CF}_n}{(1+\mathrm{IRR})^n} = 0$

Example

TWRR vs. MWRR

- Investor A and Investor B invest in a fund for two years, both use money-weighted rate of return

	Year 1	Year 2
Fund Return	Positive	Negative
Investor A	8.5%	
Investor B	10%	

Which of the following is *least likely* to be an explanation for the difference of return?

A. Investor A increased the investment in the fund at the end of year 1 whereas investor B did not make any additions or withdrawals.

B. Investor B decreased the investment in the fund at the end of year 1 whereas investor A did not make any additions or withdrawals.

C. The investors invested different amounts at inception and afterward did not make any additions or withdrawals.

- **Solution: C.**

Summary

Rates and Returns

Money-weighted & time-weighted return

Other major return measures and their Applications

- Calculate and interpret major return measures and describe their appropriate uses

Other return measures

- ***Gross return*** is the return earned by an asset manager prior to deductions for management expenses, custodial fees, taxes, or any other expenses that are not directly related to the generation of returns but rather related to the management and administration of an investment.
- ***Net return*** accounts for (i.e., deducts) all managerial and administrative expenses that reduce an investor's return.
- ***Pretax nominal return*** has no adjustment has been made for taxes or inflation.
- ***After-tax nominal return*** is computed as the total return minus any allowance for taxes on dividends, interest, and realized gains.
- ***Real return*** equals nominal return adjusted for inflation.
  - $(1+r_{\text{nominal return}})=(1+r_{\text{real return}})(1+r_{\text{inflation premium}})$
  - $(1+r_{\text{nominal return}})=(1+r_{\text{risk-free}})(1+r_{\text{risk premium}})$

Other return measures

- ***Leveraged return*** is the return on the investor's own money. Leveraging a portfolio, via borrowing or futures, can amplify the portfolio's gains or losses.

$$R_L = \frac{\text{Portfolio return}}{\text{Portfolio equity}} = \frac{\left[R_p \times (V_E + V_B) - (V_B \times r_D)\right]}{V_E} = R_p + \frac{V_B}{V_E}(R_p - r_D)$$

Example

For a RMB10 million equity portfolio that generates an 6 percent total investment return, R_P , over one year and is financed 20 percent with debt at 3 percent, then the leveraged return, R_L,?

Correct Answer:

$$R_L = R_p + \frac{V_B}{V_E}(R_p - r_D) = 6\% + \frac{20\% \times 100000}{80\% \times 100000} \times (6\% - 3\%) = 6.75\%$$

Summary

Rates and Returns

Other major return measures and their Applications

Summary

Module: Rates and Returns

Interest rates and time value of money

Annualized return

Average returns

Money-weighted and time-weighted return

Other major return measures and their Applications

Module

The Time Value of Money in Finance

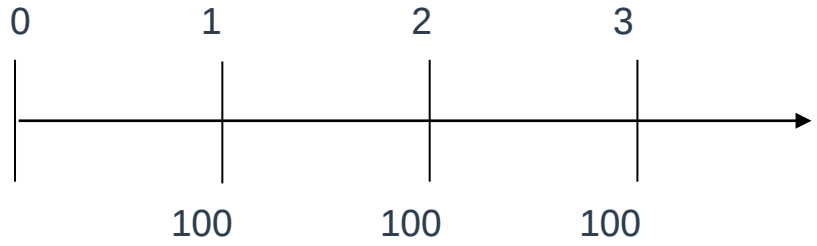
1. Annuity
2. Time value of money in fixed income and equity
3. Implied return and growth
4. Cash flow additivity

Annuity

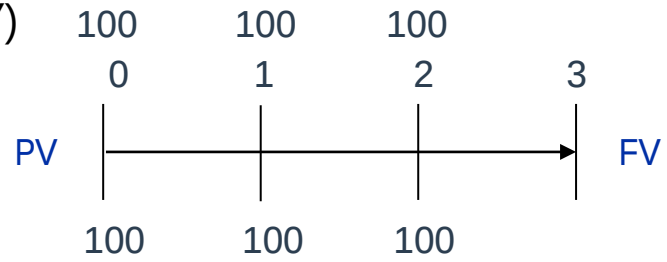
- calculate and interpret the future value (FV) and present value (PV) of a single sum of money, an ordinary annuity, an annuity due, a perpetuity (PV only), and a series of unequal cash flows
- demonstrate the use of a time line in modeling and solving time value of money problems

Annuity

- **What's annuities?**
  - is a finite set of level sequential cash flows.
    - Equal intervals.
    - Equal amount of cash flows.
    - Same direction.



- **Types of annuities**
  - **Ordinary annuity:** payments occur at the **end** of the period. (END mode)
    - The first cash flow occurs a period later (at t=1).
  - **Annuity due:** payments occur at the **beginning** of the period. (BGN mode)
    - The first cash flow occurs immediately (at t=0).
  - $FV_{Annuity\ due}/PV_{Annuity\ due} = FV_{ordinary}/Pv_{ordinary} \times (1+I/Y)$

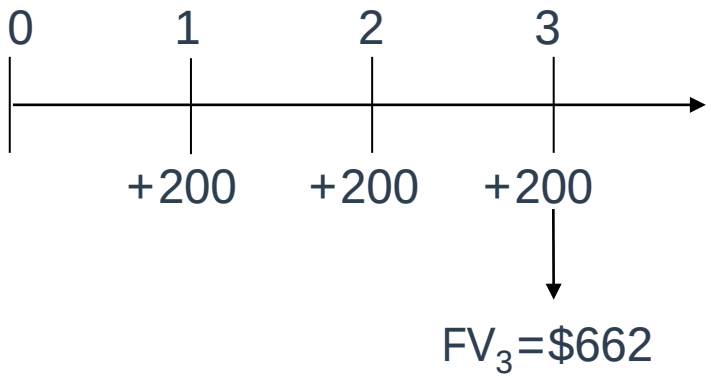


Example

Calculation of Ordinary Annuity

- What is the future value of an ordinary annuity that pays $200 per year at the end of each of the next three years, given the investment is expected to earn a 10% rate of return?

- **Correct Answer:**

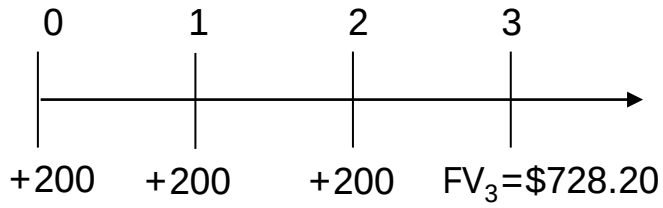


  - Enter relevant data for calculate.
    - N=3; I/Y =10; PMT=–200; PV=0; CPT→FV=$662

Example

Calculation of Annuity Due

- What is the future value of an annuity that pays $200 per year at the beginning of each of the next three years, commencing today, if the cash flows can be invested at an annual rate of 10%?



- **Correct Answer:**
  - Enter relevant data for calculate (BGN mode: ([2nd] [BGN] [2nd] [SET] [2nd] [QUIT]):
    - N=3; I/Y=10; PMT=–200; PV=0; CPT→FV_B=$728.20
  - Alternatively (END mode):
    - N=3; I/Y=10; PMT=–200; PV=0; CPT→FV_E=$662; $FV_B=FV_E\times(1+I/Y)=662\times1.10$=$728.20

Example

Amortization table

- To see how a lump sum can generate an annuity, assume that we loan $3,170 from the bank today at 10 percent interest. Construct an amortization table to show the annuity payments over the next four years.

- **Correct Answer:**

  The amount of the annuity payments: N=4; I/Y=10; PV=-$3,170; FV=0; CPT: PMT=$1,000

Amortization Table					
Time Period	Beginning Balance (1)	Payment (2)	Interest Component (3)=(1)*10%	Principal Component (4)=(2)-(3)	Ending Balance (5)=(1)-(4)
1	3,170	1000	317	683	2,487
2	2,487	1000	248.7	751.3	1,735.7
3	1,735.7	1000	173.57	826.43	909.27
4	909.27	1000	90.93	909.27	0

Example

Ordinary annuity & Annuity due

- A client plans to send a child to college for four years starting 18 years from now. Having set aside money for tuition, she decides to plan for room and board also. She estimates these costs at $20,000 per year, payable at the beginning of each year, by the time her child goes to college. If she starts next year and makes 18 payments into a saving account paying 5 percent annually, what annual payments must she make?

- **Correct Answers:**
  - Compute PV at t=18 and Set your calculator to the **BGN mode**.
    - N=4; I/Y=5; FV=0; PMT=-20,000; CPT (PV)=74,465
  - At t=18, PV18 =FV (for ordinary annuity), Set your calculator to the **END mode.**
    - N=18; I/Y=5; PV=0; FV=-74,465; CPT (PMT)=2,647.

Perpetuity

- About **perpetuity**
  - A perpetuity is a set of level never-ending sequential cash flows, with the first cash flow occurring one period from now.

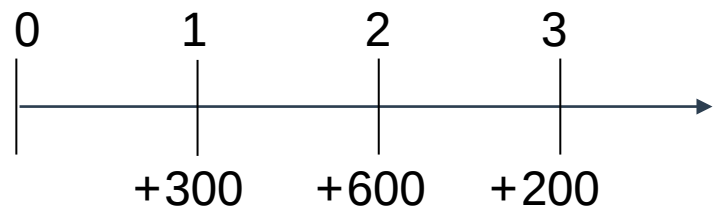
$$PV = \frac{A}{1+r} + \frac{A}{(1+r)^2} + \frac{A}{(1+r)^3} + \cdots \qquad (1)$$

$$(1+r)PV = A + \frac{A}{1+r} + \frac{A}{(1+r)^2} + \cdots \qquad (2)$$

$$(2)-(1) \qquad r \times PV = A \Rightarrow PV = \frac{A}{r}$$

Unequal Cash Flows

- About **Unequal Cash Flows**
  - A cash flow stream that is not equal from period to period.

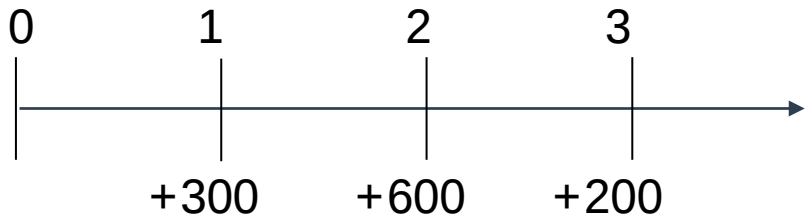


  - this series of uneven cash flows is nothing more than a stream of annual single sum cash flows.
  - To find the PV or FV of this cash flow stream, all we need to do is sum the PVs or FVs of the individual cash flows.

Example

Unequal Cash Flows

- Using a rate of return of 10%, compute the future value at the end of the third year and the present value of this three-year uneven cash flow stream described below.



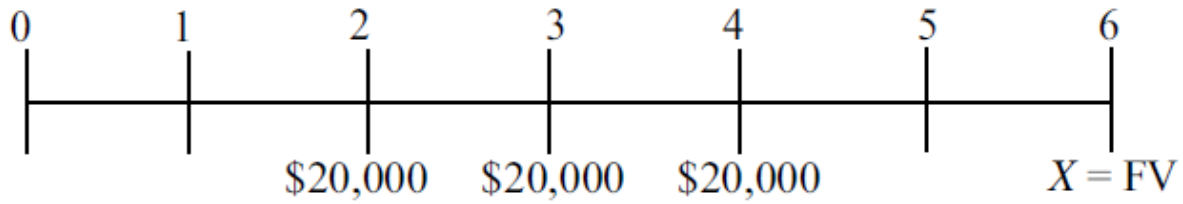
- **Correct Answer:**
  - FV1: PV = –300; I/Y = 10; N = 2; CPT → FV = FV1 = 363
    - FV2: PV = –600; I/Y = 10; N = 1; CPT → FV = FV2 = 660
    - FV3: PV = –200; I/Y = 10; N = 0; CPT → FV = FV3 = 200
    - FV of cash flow stream = $\Sigma FV_{individual}$ = 1,223
  - PV1: FV = 300; I/Y = 10; N = 1; CPT → PV = PV1 = –272.73
    - PV2: FV = 600; I/Y = 10; N = 2; CPT → PV = PV2 = –495.87
    - PV3: FV = 200; I/Y = 10; N = 3; CPT → PV = PV3 = –150.26
    - PV of cash flow stream = $\Sigma PV_{individual}$ = $918.86

Example: Annuity

- Two years from now, a client will receive the first of three annual payments of $20,000 from a small business project. If she can earn 9 percent annually on her investments and plans to retire in six years, how much will the three business project payments be worth at the time of her retirement?

Example: Annuity

- **Correct Answer:**
  - **i.** Draw a time line.



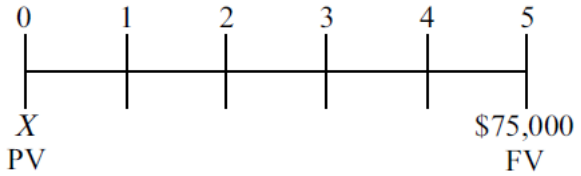
  - **ii.** Recognize the problem as the future value of a delayed annuity. Delaying the payments requires two calculations.
  - **iii.** To bring the three $20,000 payments to an equivalent lump sum of $65,562.00 four years from today:
    - FV_4: PV = 0; I/Y = 9; N = 3; PMT=-20,000 CPT → FV_4 = 65562
  - **iv.** Then use the formula for the future value of a lump sum, $FV_N = PV(1 + r)^N$,to bring the single lump sum of $65,562.00 to an equivalent lump sum of $77,894.21 six years from today.

Example: Annuity

- To cover the first year's total college tuition payments for his two children, a father will make a $75,000 payment five years from now. How much will he need to invest today to meet his first tuition goal if the investment earns 6 percent annually?

Example: Annuity

- **Correct Answer:**
  - i. Draw a time line.



  - ii. Identify the problem as the present value of a lump sum.
  - iii. Use the formula for the present value of a lump sum.

$$\text{PV} = FV_N(1 + r)^{-N} = 75000(1 + 0.06)^{-5} = 56044.36$$

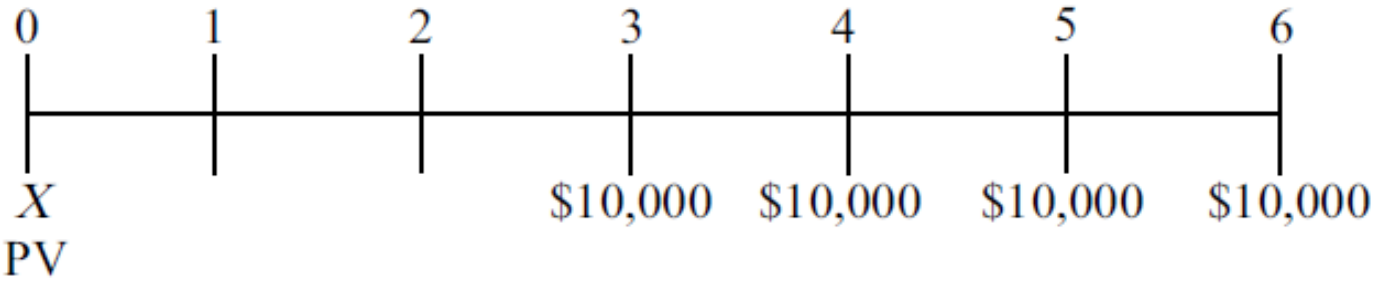
In summary, the father will need to invest $56,044.36 today in order to have $75,000 in five years if his investments earn 6 percent annually.

Example: Annuity

- Suppose you plan to send your daughter to college in three years. You expect her to earn two-thirds of her tuition payment in scholarship money, so you estimate that your payments will be $10,000 a year for four years. To estimate whether you have set aside enough money, you ignore possible inflation in tuition payments and assume that you can earn 8 percent annually on your investments. How much should you set aside now to cover these payments?

Example: Annuity

- **Correct Answer:**
  - **i.** Draw a time line.



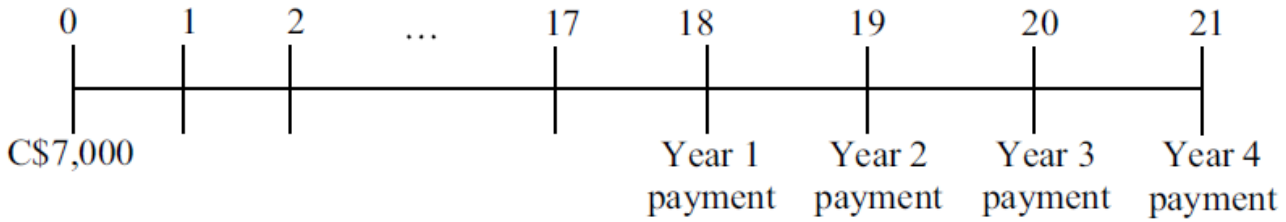
  - **ii.** Recognize the problem as a delayed annuity. Delaying the payments requires two calculations.
  - **iii.** Giving an ordinary annuity: 3时刻折现回2时刻（END）
    - FV = 0; I/Y = 8; N = 4; PMT=-10,000 CPT → PV2 = 33,121.27
    - PV = PV2 $(1 + 8\%)^{-2}$= $28,396.15
  - **iii.** Giving an annuity due: 3时刻折现回3时刻（BGN）
    - FV = 0; I/Y = 8; N = 4; PMT=-10,000 CPT → PV3 = 35,770.97
    - PV = PV3 $(1 + 8\%)^{-3}$= $28,396.15

Example: Annuity

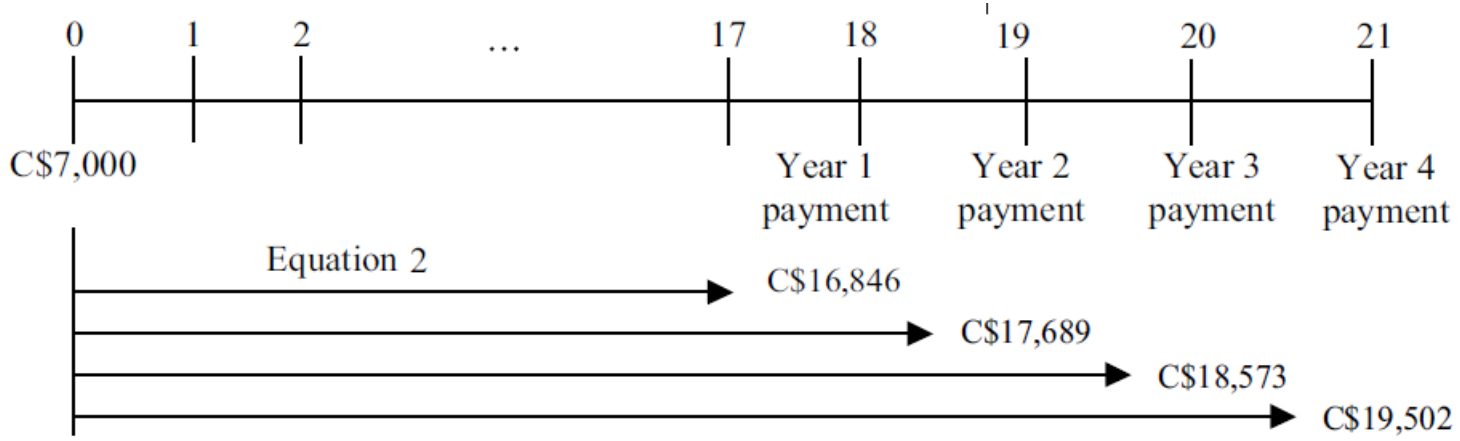
- A couple plans to pay their child's college tuition for 4 years starting 18 years from now. The current annual cost of college is C$7,000, and they expect this cost to rise at an annual rate of 5 percent. In their planning, they assume that they can earn 6 percent annually. How much must they put aside each year, starting next year, if they plan to make 17 equal payments?

Example: Annuity

- **Correct Answer:**
  - **i.** Draw a time line.



  - **ii.** Recognize that the payments in Years 18, 19, 20, and 21 are the future values of a lump sum of C$7,000 in Year 0.
  - **iii.** With r = 5%, use the formula for the future value of a lump sum, FVN =PV$(1 + r)^N$, four times to find the payments. These future values are shown on the time line below.



(continued)

Example: Annuity

- **Correct Answer:**
  - **iv**. Using the formula for the present value of a lump sum (r = 6%), equate the four college payments to single payments as of t = 17 and add them together.
    - $16{,}846(1.06)^{-1} + 17{,}689(1.06)^{-2} + 18{,}573(1.06)^{-3} + 19{,}502(1.06)^{-4} = 62{,}677$
  - **v.** Then use Calculator:
    - FV = 62,677; I/Y = 6; N = 17; PV=0 CPT → PMT = 2,221.58

Summary

The Time Value of Money in Finance

Annuity

Calculation of ordinary annuity

Calculation of annuity due

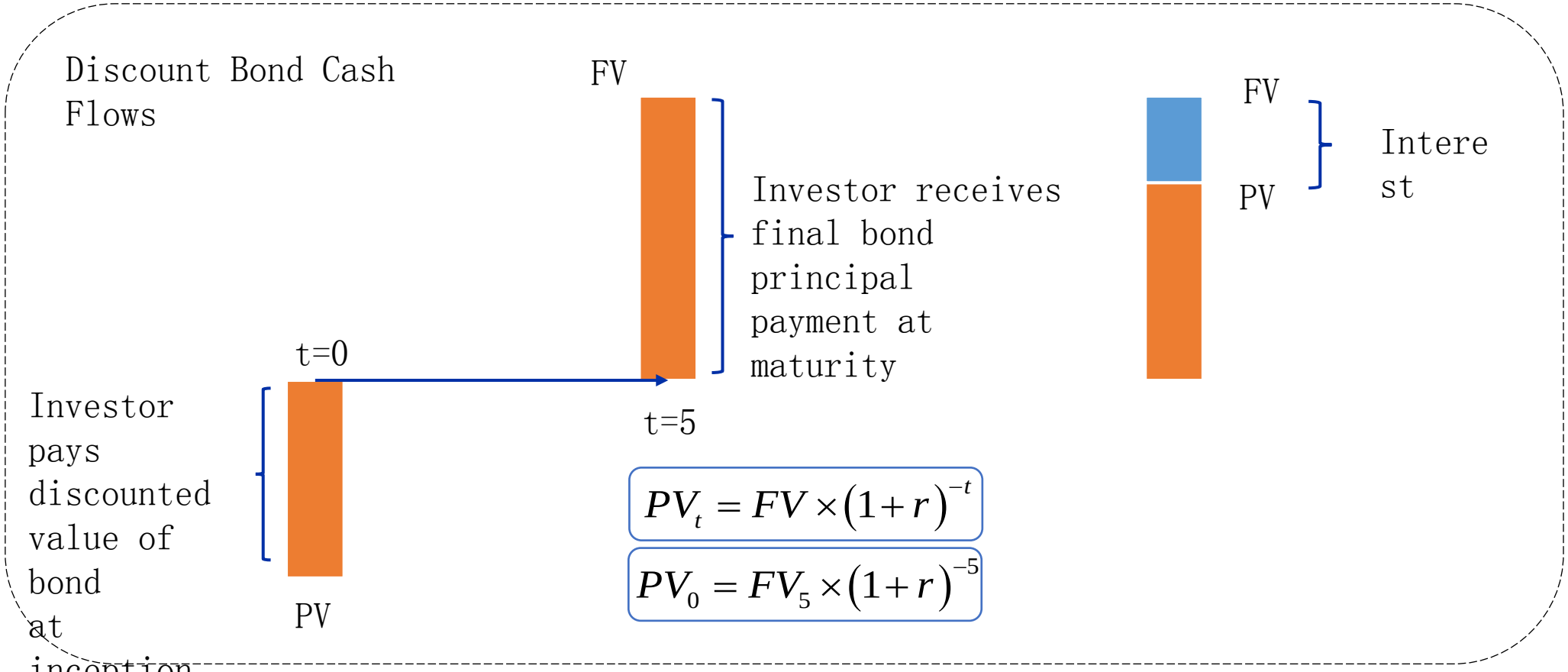
Application of other annuities

Time value of money in fixed income and equity

- Calculate and interpret the present value (PV) of fixed-income and equity instruments based on expected future cash flows

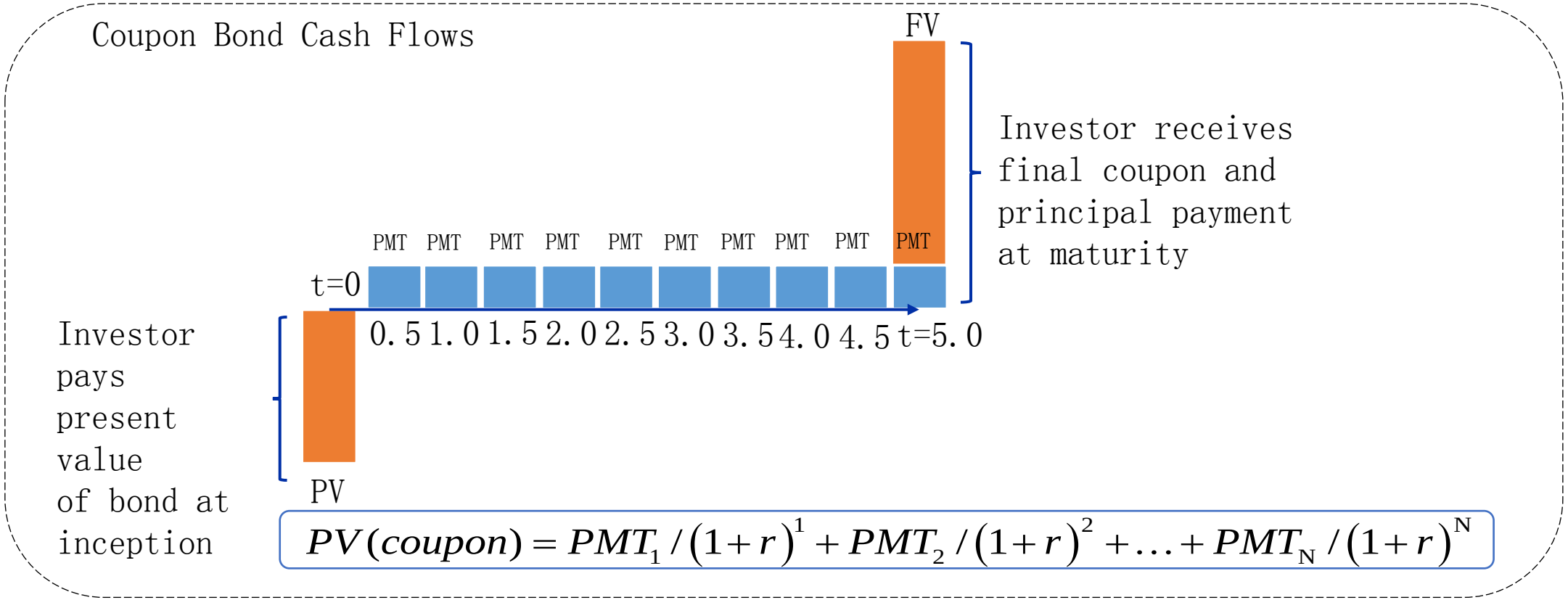
Fixed-Income Instruments

- **Discount bond (zero-coupon bond)**
  - An investor pays an initial price (PV) for a bond or loan and receives a single principal cash flow (FV) at maturity. The difference (FV − PV) represents the interest earned over the life of the instrument.



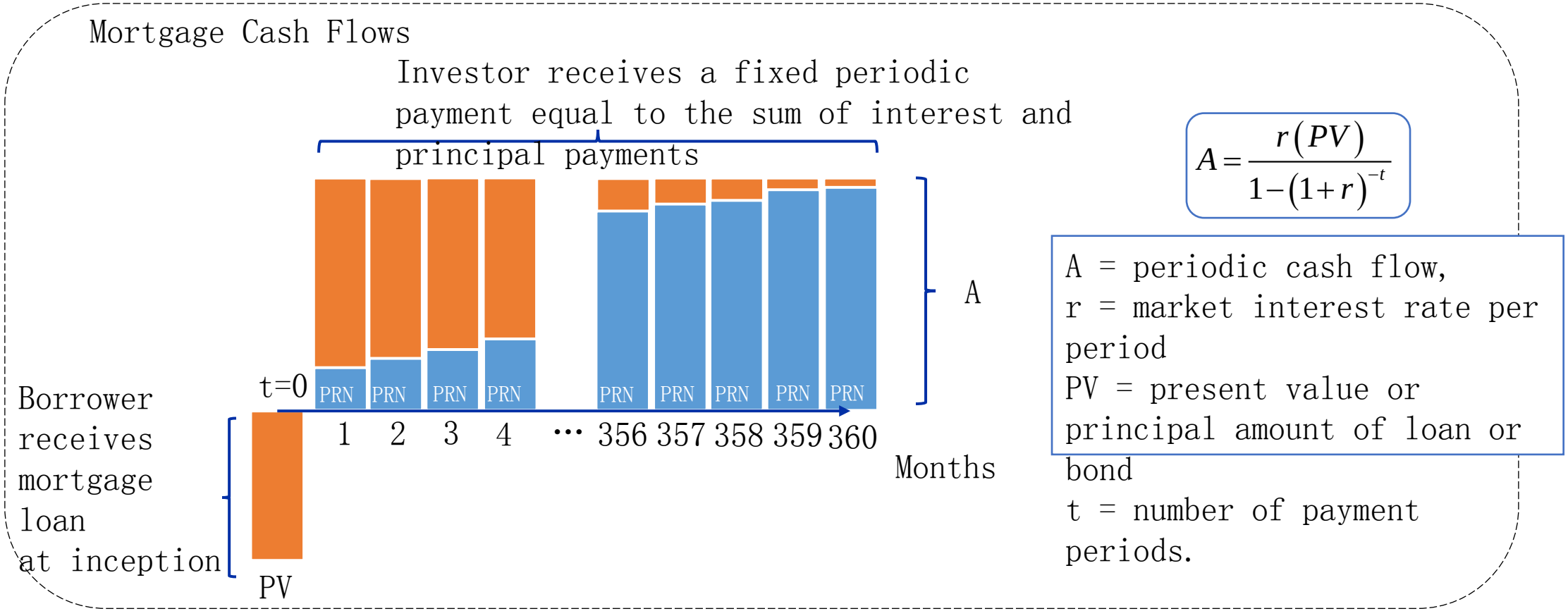
Fixed-Income Instruments

- **Periodic Interest**
  - An investor pays an initial price (PV) for a bond or loan and receives interest cash flows (PMT) at pre-determined intervals over the life of the instrument, with the final interest payment and the principal (FV) paid at maturity.



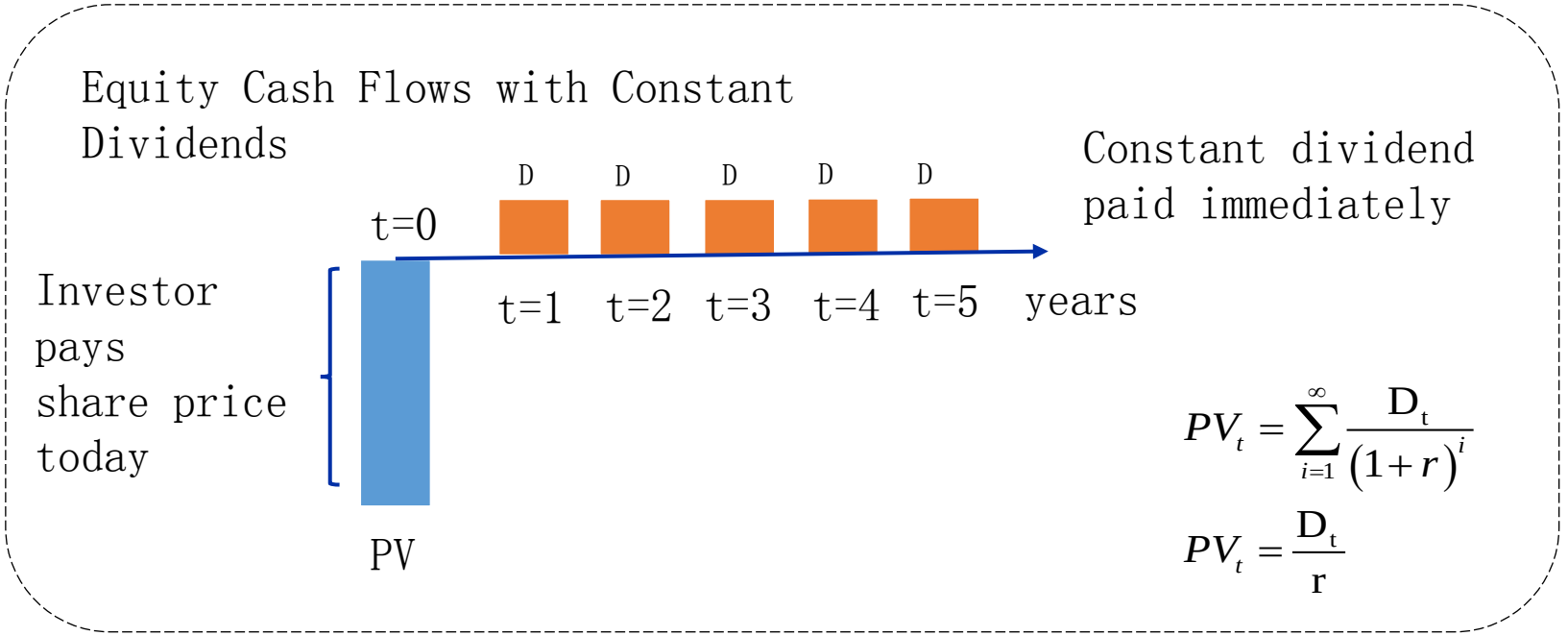
Fixed-Income Instruments

- **Level Payments (fully amortizing loans)**
  - An investor pays an initial price (PV) and receives uniform cash flows at pre-determined intervals (A) through maturity which represent both interest and principal repayment.



Equity Instruments

- **Constant Dividends**
  - An investor pays an initial price (PV) for a preferred or common share of stock and receives a fixed periodic dividend (D).

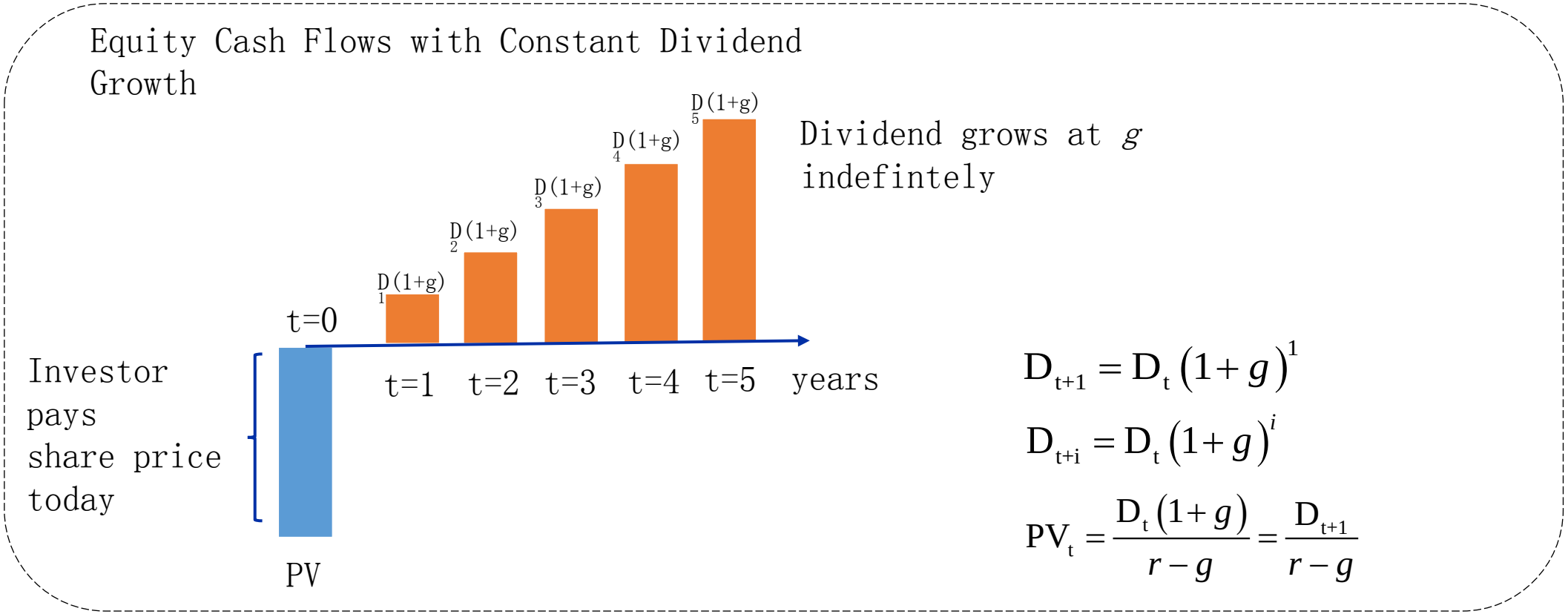


$$PV_t = \sum_{i=1}^{\infty} \frac{D_t}{(1+r)^i}$$

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

Equity Instruments

- **Constant Dividend Growth Rate:**
  - An investor pays an initial price (PV) for a share of stock and receives an initial dividend in one period (D_{t+1}), which is expected to grow over time at a constant rate of g.



$$D_{t+1} = D_t (1+g)^1$$

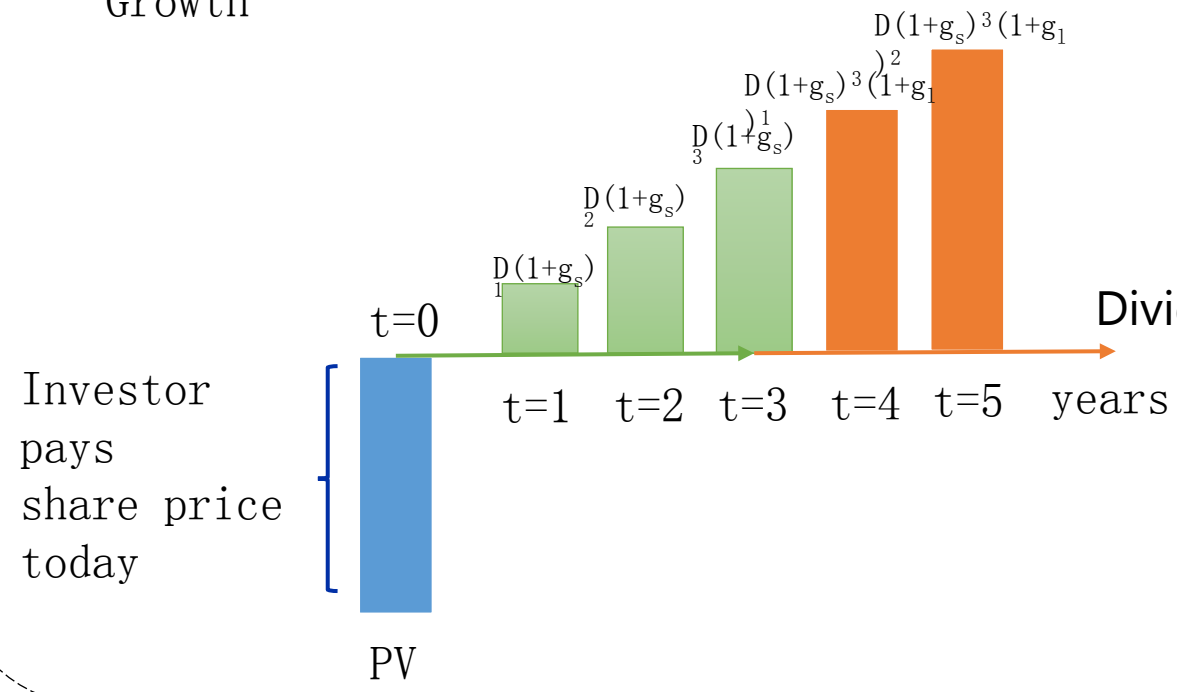
$$D_{t+i} = D_t (1+g)^i$$

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

Equity Instruments

- **Changing Dividend Growth Rate:**
  - An investor pays an initial price (PV) for a share of stock and receives an initial dividend in one period (Dt+1). The dividend is expected to grow at a rate that changes over time as a company moves from an initial period of high growth to slower growth as it reaches maturity.

Equity Cash Flows with Two-Stage Dividend Growth



g_s: Short term growth rate

g_l: Long term growth rate

Dividend grows at g_l indefinitely

$$PV_t = \sum_{i=1}^{n} \frac{D_t(1+g_s)^i}{(1+r)^i} + \sum_{j=n+1}^{\infty} \frac{D_{t+n}(1+g_l)^i}{(1+r)^j}$$

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

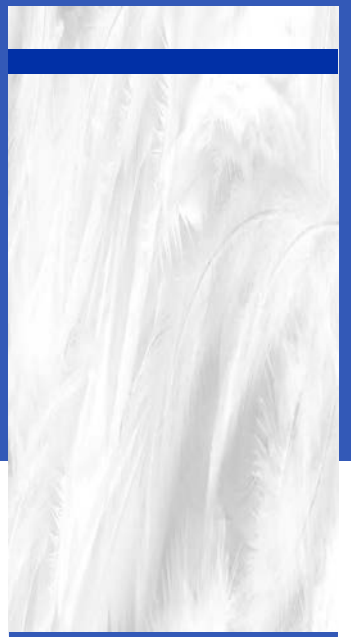
Summary

The Time Value of Money in Finance

Time value of money in fixed income and equity

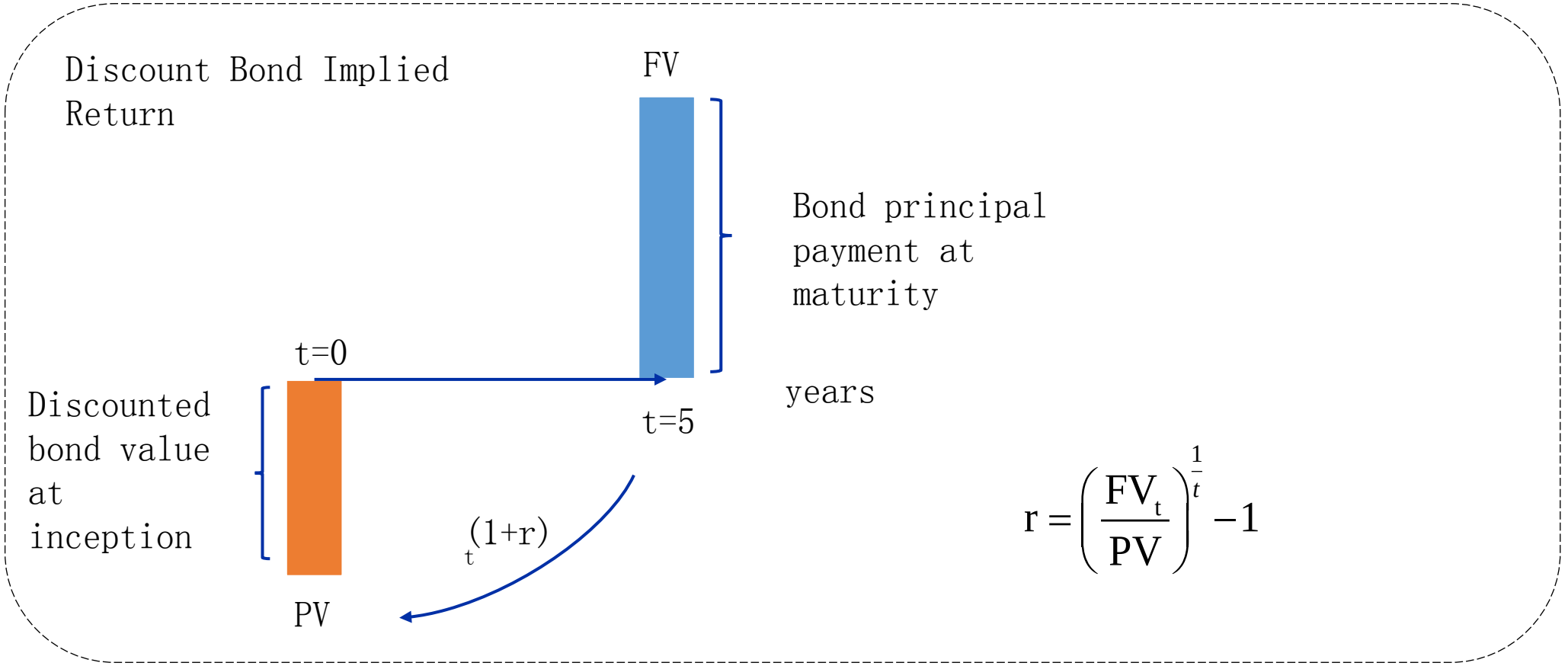
Implied return and growth

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



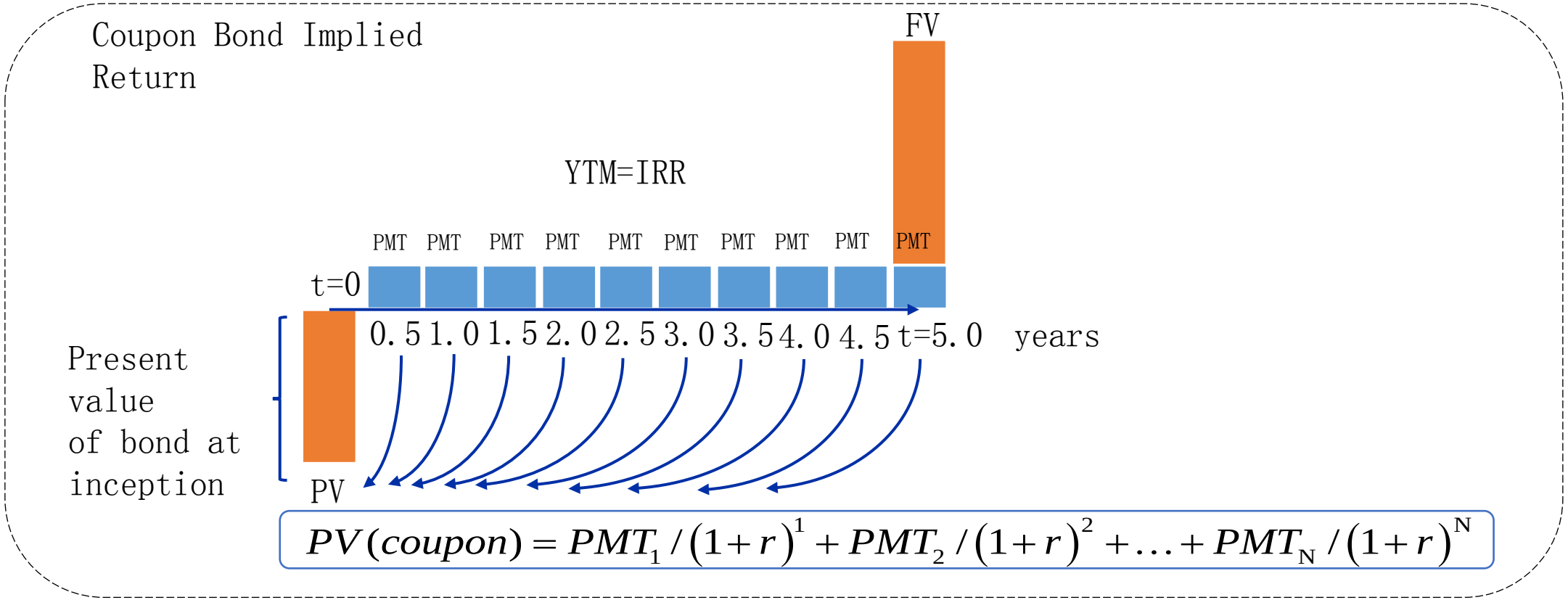
Implied Return for Fixed-Income Instruments

- **Discount Bond Implied Return**
  - Given observed present value (or price) and all future cash flows occur as promised, the discount rate (r) or yield-to-maturity (YTM) is a measure of implied return for the cash flow pattern.



Implied Return for Fixed-Income Instruments

- **Coupon Bond Implied Return**
  - The uniform discount rate (or internal rate of return) for all promised cash flows is the YTM, a single implied market discount rate for all cash flows regardless of timing.



Implied r & g for Equity Instruments

- **Implied return & implied growth for equity instruments**
  - By calculating the present value of an equity investment, the price of a share of stock reflects not only the required return but also the growth of cash flows. If we begin with an assumption of constant growth of dividends.

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

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

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

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

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

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

g: implied growth rate
r: implied return
E_t: earnings per share for period t
D_t/E_t: dividend payout ratio
D_{t+1}/E_{t+1}: expected dividend payout ratio
PV_t/E_t: price-to-earning ratio
PV_t/E_{t+1}: forward price-to-earning ratio

Example
Implied Return and Growth from Price to Earnings Ratio

- Suppose stock CC trades at a forward price to earnings ratio of 28 and its expected dividend payout ratio is 70%. Analysts believe that stock CC should earn a 9% return and that its dividends will grow by 4.50% per year indefinitely. Recommend a course of action for an investor interested in taking a position in stock CC.

- **Solution**:
  - An investor should consider a short position in stock CC in the belief that its price should decline because its PV_t/E_{t+1} (28) is well above what its fundamentals (15.56) imply and therefore PV_t/E_{t+1} is overvalued.

g: implied growth rate=4.5%
r: implied return=9%
D_{t+1}/E_{t+1}: expected dividend payout ratio=70%
PV_t/E_{t+1}: forward price-to-earning ratio=28

$$\frac{PV_t}{E_{t+1}} = \frac{\frac{D_{t+1}}{E_{t+1}}}{r-g} \qquad 28 > \frac{0.7}{0.09-0.045} = 15.56$$

Summary

The Time Value of Money in Finance

Implied return and growth

Cash flow additivity

- 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

Cash Flow Additivity

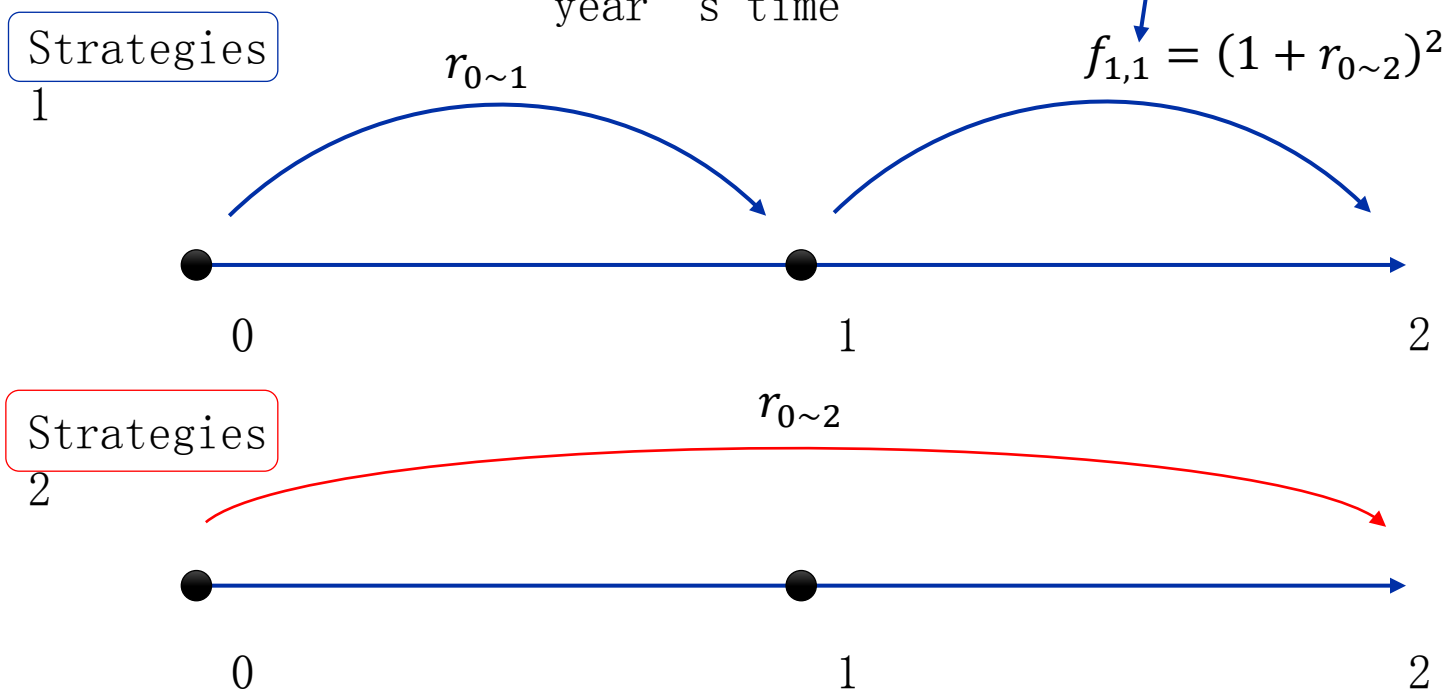
- **Cash flow additivity principle**
  - Under cash flow additivity, the present value of any future cash flow stream indexed at the same point equals the sum of the present values of the cash flows. No possibility exists to earn a riskless profit in the absence of transaction costs.
  - Three economic situations illustration.
    - Implied Forward Rates Using Cash Flow Additivity
    - Forward Exchange Rates Using No Arbitrage
    - Option Pricing Using Cash Flow Additivity

Implied Forward Rates

- **Implied Forward Rates Using Cash Flow Additivity**
  - Under the cash flow additivity principle, a risk-neutral investor would be indifferent between strategies 1 and 2.

Implied Forward Rates or
the breakeven one-year reinvestment rate in one year's time

$$f_{1,1} = (1 + r_{0\sim2})^2 \div (1 + r_{0\sim1})^1 - 1$$

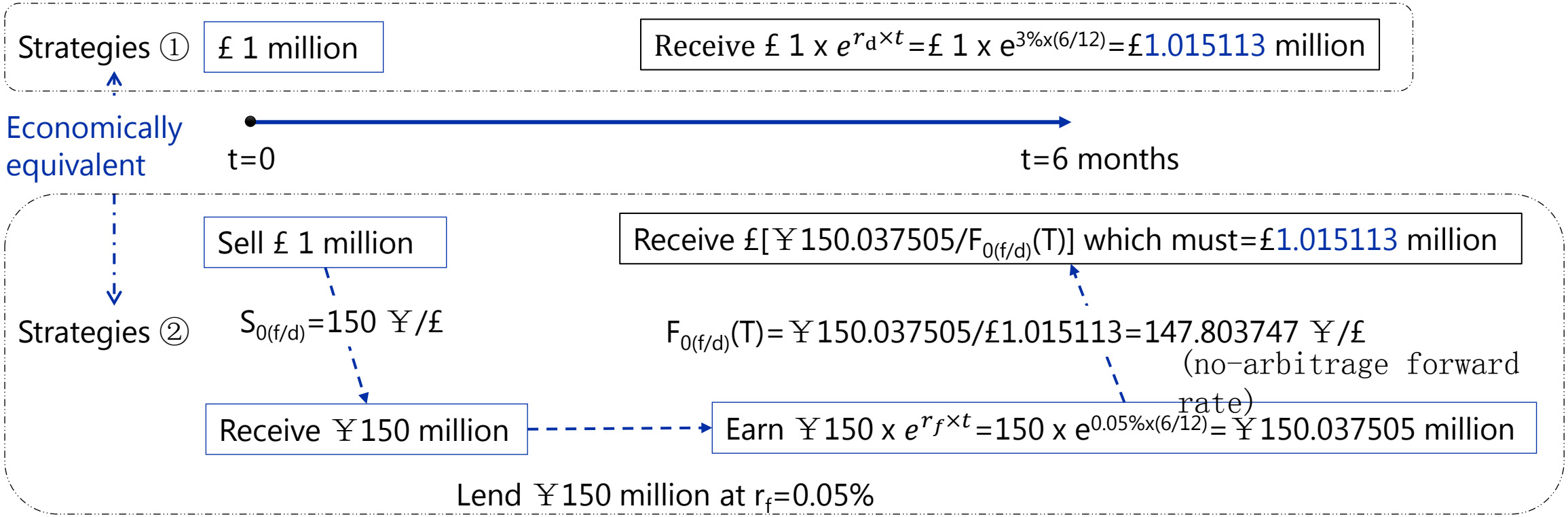


$$FV_2 = PV_0(1+ r_{0\sim1})^1(1+ f_{1,1})^1 = PV_0(1+ r_{0\sim2})^2$$

$$(1+ r_{0\sim1})^1(1+ f_{1,1})^1 = (1+ r_{0\sim2})^2$$

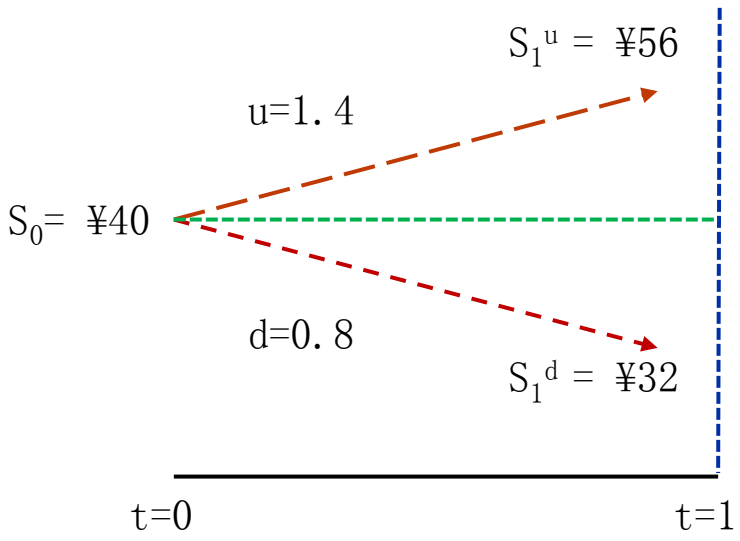
Forward Exchange Rates

- **No-arbitrage forward rate**
  - An investor has £1 million to make a riskless investment in either British or Japanese six-month government debt for six months. The current exchange rate between JPY and GBP is 150 JPY/GBP. The six-month Japanese yen r_f is 0.05%, and the six-month British pound r_d is 3%. (continuous compounding).

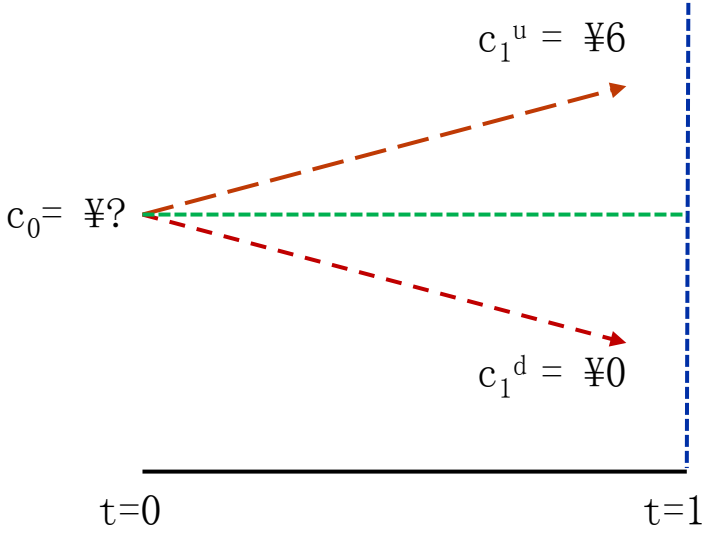


Option Pricing

- **Option pricing using cash flow activity and no-arbitrage pricing**
  - A stock's current price (S_0) is ¥40. The price may rise 40% to ¥56 or may fall 20% to ¥32 during the next time period(t=1). An investor wishes to sell a contract on the stock (short call option position), in which the buyer of the contract has the right, but not obligation, to buy the noted stock(underlying asset) for ¥50 (exercise price X) at t=1. To establish no-arbitrage pricing for this contract(call option).



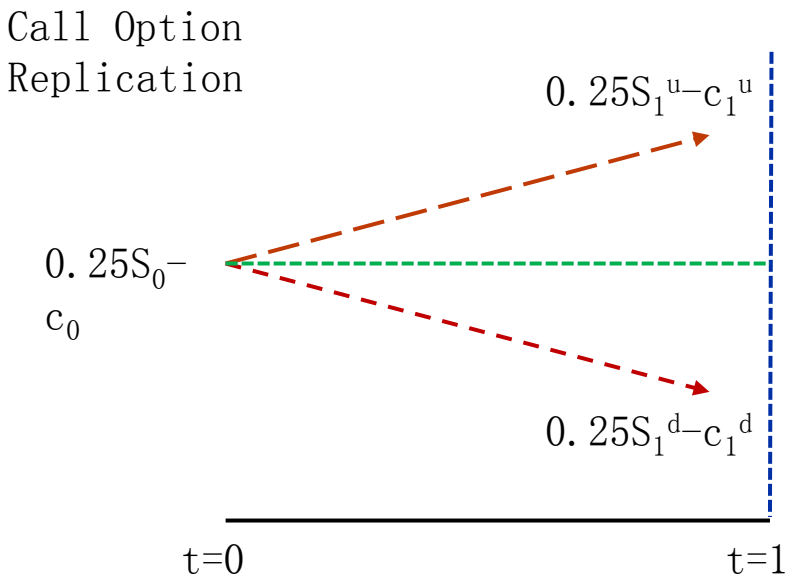
One-Period Binomial Tree for the stock's Price



One-Period Binomial Tree for the Contract's Price

Option Pricing

- **Option pricing using cash flow activity and no-arbitrage pricing**



+S + −c → △portfolio's value $_{t=1}=V_1=0$

+hS + −c = 0

+ h△S + − △ c = 0

h = △c/△S = (6-0)/(56-32)= 0.25 = Hedge ratio

$V_1=0.25S_1^u-c_1^u=0.25S_1^d-c_1^d$

$V_0=0.25S_0-c_0$

$V_0=V_1$

$V_0=0.25S_0-c_0=(0.25S_1^u-c_1^u)/(1+R^f)^1=(0.25S_1^d-c_1^d)/(1+R^f)^1=V_1$

$c_0=0.25x40-8/(1+R_f)^1$

Summary

The Time Value of Money in Finance

Cash flow additivity

Summary

Module: The Time Value of Money in Finance

Annuity

Time value of money in fixed income and equity

Implied return and growth

Cash flow additivity

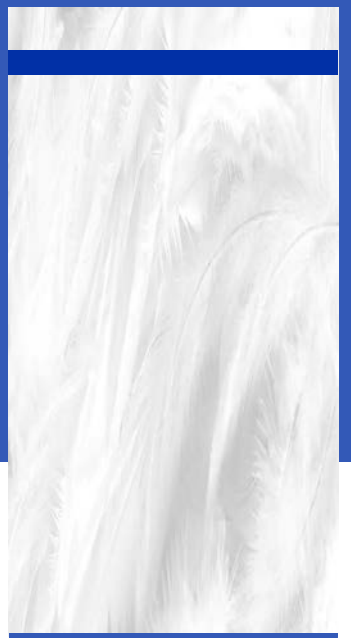
Module

Statistical Measures of Asset Returns

1. Measures of Central Tendency and Location
2. Measures of Dispersion
3. Measures of Shape of a Distribution
4. Correlation between Two Variables

Measures of central tendency and location

- Calculate, interpret, and evaluate measures of central tendency and location to address an investment problem



Measures of central tendency

- **Arithmetic mean** is the sum of the values of the observations in a dataset divided by the number of observations.

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

  - E.g., the sample mean or average, $\overline{X}$ (read "X-bar"), is the arithmetic mean value of a sample.
  - All observations used; equal weights; sensitive to extreme values (outliers); mathematically tractable.
- **Median is** the value of the middle item of a set of items that has been sorted into ascending or descending order.
  - In an odd-numbered sample of n items, the median is the value of the item that occupies the (n + 1)/2 position.
  - In an even-numbered sample, the median is the mean of the values of items occupying the n/2 and (n + 2)/2 positions (the two middle items).
  - A distribution has only one median; outliers do not affect median; `calculation is complex`.

Measures of central tendency

- **The mode** is the most frequently occurring value in a dataset.
  - A dataset can have more than one mode, or even no mode.
    - Unimodal distribution: a dataset has a single value that is observed most frequently;
    - Bimodal distribution: a dataset has two most frequently occurring values, then it has two modes;
    - No mode: when all the values in a dataset are different, no value occurs more frequently than any other value.

Dealing with Outliers

- Representing a rare value in the population, an outlier may reflect an error in recording the value of an observation or an observation generated from a different population.
  - Option 1 **Do nothing**; use the data (contain meaningful information) without any adjustment.
  - Option 2 **Delete** all the outliers; use trimmed mean;
  - Option 3 **Replace** the outliers with another value, winsorized mean.
- Application: to reveal important insights about the dataset, e.g., analyzing the behavior of asset returns and rate, price, spread and volume changes, by comparing the statistical measures of datasets with outliers included and with outliers excluded.

Example

Handling Outliers: Daily Returns to an Index

- Using daily returns on a Equity Index for the past five years, consisting of 1,258 trading days, the effect of trimming and winsorizing the data is shown below:

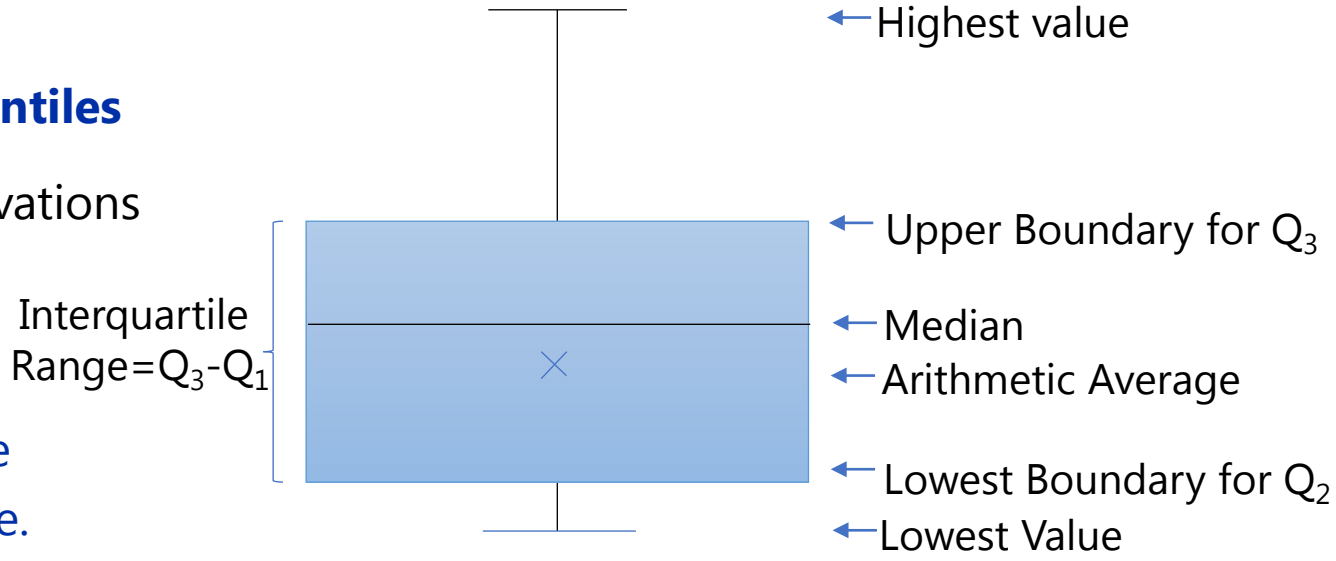
Effect of Trimming and Winsorizing

	Arithmetic Mean	Trimmed Mean (Trimmed 5%)	Winsorized Mean (95%)
Mean	0.04%	0.05%	0.04%
Number of Observations	1,258	1,194	1,258

- The trimmed mean eliminates the lowest 2.5 percent of returns, which in this sample is any daily return less than −1.934 percent, and it eliminates the highest 2.5 percent, which in this sample is any daily return greater than 1.671 percent. The result of this trimming is that the mean is calculated using 1,194 observations instead of the original sample's 1,258 observations.
- The winsorized mean substitutes a return of −1.934 percent (the 2.5 percentile value) for any observation below −1.934 and substitutes a return of 1.671 percent (the 97.5 percentile value) for any observation above 1.671.
- The trimmed and winsorized means are higher than the arithmetic mean, suggesting the potential evidence of significant negative returns in the observed daily return distribution.

Measures of Location

- **Quantiles: quartiles/quintiles/deciles/percentiles**
  - The third quintile: there are 60% the observations fall at or below that value.
- **Calculation: $L_y = (n+1) * y/100$.**
- The **interquartile range** (IQR) is the difference between the third quartile and the first quartile.



Example

Observers in ascending order: 5 8 11 12 14 16 16 18 19 21 23, **calculate the third quartile.**

Correct Answer:

N=11 , L_y=(11+1)*75%=9, i.e. the 9th number is 75%

The third quartiles = 19

Example

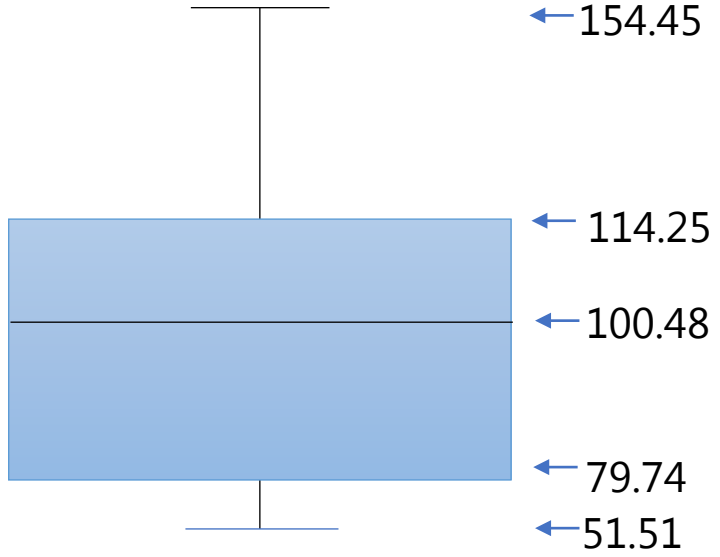
Quantiles

- Consider the box and whisker plot and answer two questions.
  - The median and the interquartile range are closest to:

A. 34.51 and 13.76.

B. 100.48 and 34.51.

C. 102.98 and 25.74.



- **Correct Answers: B.**
  - The median is indicated within the box, which is 100.48.
  - The interquartile range is the height of the box, which is the difference between 114.25 and 79.74, equal to 34.51.

Summary

Statistical Measures of Asset Returns

Measures of Central Tendency and Location

Measures of Central Tendency

Dealing with Outliers

Measures of Location

Measures of dispersion

- Calculate, interpret, and evaluate measures of dispersion to address an investment problem

Absolute Dispersion

- **Absolute dispersion** is the amount of variability present **without** comparison to any reference point or benchmark.
  - **Range = maximum value – minimum value**
  - **Mean absolute deviation (MAD)** $= \frac{\sum_{i=1}^{N} |X_i - \bar{X}|}{n}$
  - **Variance (Var) & standard deviation (S.D.)**
    - **For population:** $\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$ $\sigma = \sqrt{\sigma^2}$
    - **For sample:** $s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$ $s = \sqrt{s^2}$
  - **Downside deviation**

$$\text{Semivariance} = \frac{\sum_{\text{for all } X_i \le \bar{X}} (X_i - \bar{X})^2}{n-1} \qquad \text{Target Semivariance} = \frac{\sum_{\text{for all } X_i \le B} (X_i - B)^2}{n-1}$$

- Relationship between the arithmetic mean and the geometric mean
  - $G \approx A - S^2/2$
- The more disperse or volatile the returns, the larger the gap between the geometric mean return and the arithmetic mean return.

Relative dispersion

- **Relative dispersion** is the amount of dispersion relative to a reference value or benchmark.
  - **Coefficient of variation** measures the amount of dispersion in a distribution relative to the Arithmetic mean.

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

    - Relative dispersion.
    - Scale free.

  - The **Sharpe ratio** measures excess return per unit of risk.

$$\text{Sharpe ratio} = \frac{E(R_p) - R_f}{\sigma_p}$$

Example

Absolute Dispersion

- As a mutual fund analyst, you are examining, as of early 2017, the most recent five years of total returns for two US large-cap value equity mutual fund.

Year	Portfolio return(%)
2012	-39.44
2013	31.64
2014	12.53
2015	-4.35
2016	12.82

The portfolio's mean absolute deviation and variance of annual returns, respectively, for the five-year period are closest to:

	Mean absolute deviation	Population Variance
A.	19.63%	0.05724
B.	2.64%	0.0968
C.	19.63%	0.0968

- **Correct Answer: A**

Example

Calculating Absolute Dispersion

- **Calculation process (cont.)**

输入统计数据 （金融计算器）	
按下2nd 7 (DATA)	屏幕上显示X01及其先前的值
按下2nd CLR WORK	清空工作表
键入X01的值，按下enter键	输入第一个变量值
按↓键，屏幕显示Y01	默认值为1
再按↓键，显示下一个X变量	重复第三至四步，直到把所有X变量输完为止

查找统计结果	
按下2nd 8 (STAT)	LIN
按↓键，屏幕显示n	样本量
按↓键，屏幕显示	变量X的均值
按↓键，屏幕显示S_x	变量X的样本标准差
按↓键，屏幕显示σ_x	变量X的总体标准差

Example 1: Calculating Sample Standard Deviation

- Given the data in Exhibit 1 below, calculate the sample standard deviation.

Monthly Portfolio Returns

Month	Return (%)
January	5
February	3
March	−1
April	−4
May	4
June	2
July	0
August	4
September	3
October	0
November	6
December	5

Example 1: Calculating Sample Standard Deviation

- **Correct Answer:**
  - The sample standard deviation is $\sqrt{96.2500/(12-1)}$=2.958%

Month	Observation	Deviation from the mean	Squared deviation
January	5	2.75	7.5625
February	3	0.75	0.5625
March	−1	−3.25	10.5625
April	−4	−6.25	39.0625
May	4	1.75	3.0625
June	2	−0.25	0.0625
July	0	−2.25	5.0625
August	4	1.75	3.0625
September	3	0.75	0.5625
October	0	−2.25	5.0625
November	6	3.75	14.0625
December	5	2.75	7.5625
Sum	27		**96.2500**

Example 2: Calculating Target Downside Deviation

- Based on the data in previous question and Exhibit 1, calculate the target downside deviation (target=2%).
- **Correct Answer:** The target semi-deviation with 2% target = $\sqrt{53/(12-1)}=2.195\%$

Month	Observation	Deviation from the 2% Target	Deviations below the Target	Squared Deviations below the Target
January	5	3	—	—
February	3	1	—	—
March	−1	−3	−3	9
April	−4	−6	−6	36
May	4	2	—	—
June	2	0	—	—
July	0	−2	−2	4
August	4	2	—	—
September	3	1	—	—
October	0	−2	−2	4
November	6	4	—	—
December	5	3	—	—
Sum				**53**

Example
Comparing the Target Downside Deviation with the S.D.

- Compare the standard deviation, the target downside deviation if the target is 2%, and the target downside deviation if the target is 3%

- **Correct Answer:**
  - The standard deviation is based on the deviation from the mean, which is 2.25%.The standard deviation includes all deviations from the mean, not just those below it. This results in a sample standard deviation of 2.958%.
  - Considering just the four observations below the 2% target, the target semi-deviation is 2.195%. It is **less than** the sample standard deviation since target semi-deviation captures only the downside risk (i.e., deviations below the target).
  - Considering target semi-deviation with a 3% target, there are now five (+July) observations below 3%, so the target semi-deviation is **higher**, at 2.763%.

Summary

Statistical Measures of Asset Returns

Measures of dispersion

The Range

Mean Absolute Deviations

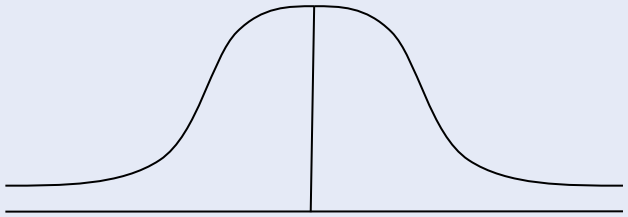
Sample Variance and Sample Standard Deviation

Downside Deviation and Coefficient of Variation

Measures of shape of a distribution

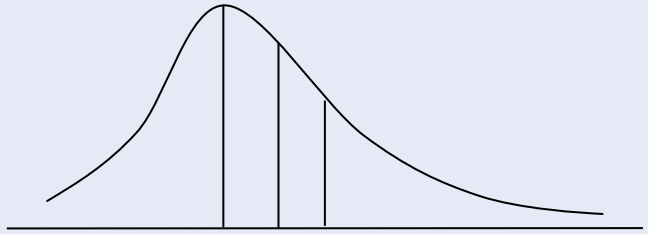
- interpret and evaluate measures of skewness and kurtosis to address an investment problem

Skewness



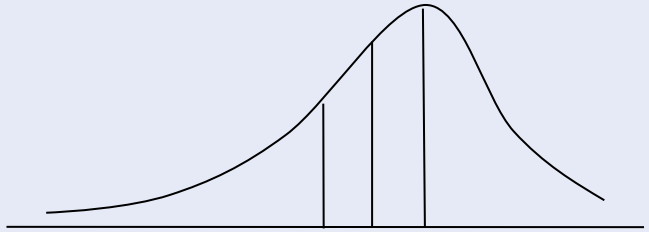
Mean=Median=Mode

Symmetrical



Mode<Median<Mean

Positive (right) skew



Mean<Median<Mode

Negative (left) skew

- A distribution that is not symmetrical is termed **skewed.**
  - **Positively skewed**: Mode<median<**mean**, having a long tail on the **right** side.
    - A return distribution with positive skew has frequent small losses and few extreme gains.
  - **Negatively skewed**: Mode>median>**mean**, having a long tail on the **left** side.
    - A return distribution with negative skew has frequent small gains and few extreme losses.
- Investors favor **a positively skewed returns** because the mean return falls above the median.
- **Sample skewness:**

$$S_K = \left[\frac{n}{(n-1)(n-2)}\right]\frac{\sum_{i=1}^{n}(X_i-\overline{X})^3}{s^3} \approx \left(\frac{1}{n}\right)\frac{\sum_{i=1}^{n}(X_i-\overline{X})^3}{s^3}$$

the third power

Example

- Is a return distribution characterized by frequent small losses and a few large gains best described as having:

	Negative skew?	A mean that is greater than the median?
A	No	No
B	No	Yes
C	Yes	No

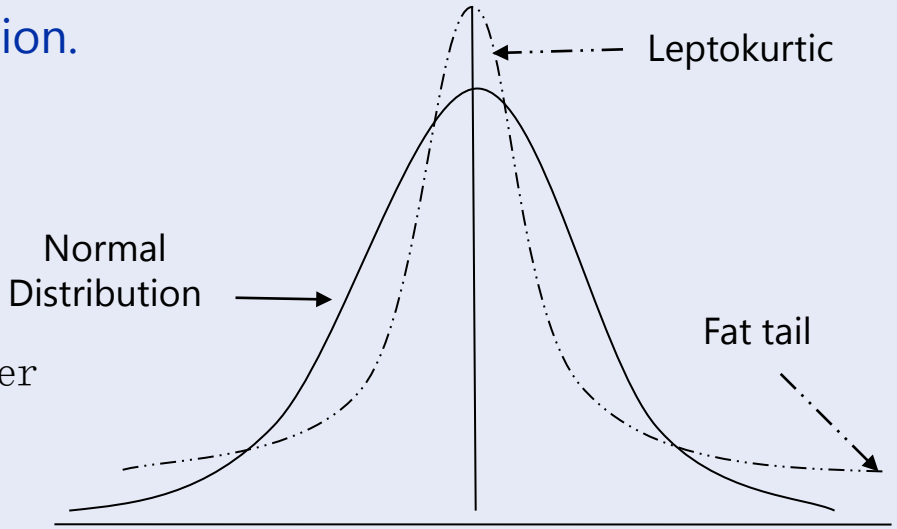
- **Correct Answers: B.**

  A distribution with frequent small losses and a few large gains is positively skewed (long tail on the right side) and the mean is greater than the median.

Kurtosis

- **Kurtosis** is a measure of the combined weight of the tails of a distribution relative to the rest of the distribution, e.g., the proportion of the total probability that is outside of, say, 2.5σ of the mean.
- Comparison with a normal distribution (kurtosis=3. 0)
  - Mesokurtic: a distribution similar to the normal distribution as it concerns relative weight in the tails;
  - Leptokurtic (fat-tailed): a distribution with fatter tails;
  - platykurtic (thin-tailed): a distribution with thinner tails.
- **Excess kurtosis** is the kurtosis relative to the normal distribution.

	Leptokurtic	Normal distribution	Platykurtic
Sample kurtosis	>3	=3	<3
Excess kurtosis	>0	=0	<0



- **Sample kurtosis and Excess kurtosis**

the fourth power

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{s^4} \approx \frac{1}{n} \frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{s^4}$$

$$K_E = K - 3$$

Example

- An analyst gathered the following information about the return distribution for two portfolios during the same time period:

Portfolio	Skewness	Kurtosis
A	-1	2.8
B	0.8	4.2

The analyst stated that the distribution for Portfolio A is less peaked than a normal distribution and that the distribution for Portfolio B has a long tail on the left side of the distribution. Is the analyst's statement correct with respect to:

	Portfolio A	Portfolio B
A.	No	No
B.	No	Yes
C.	Yes	No

- **Correct Solution: C.**

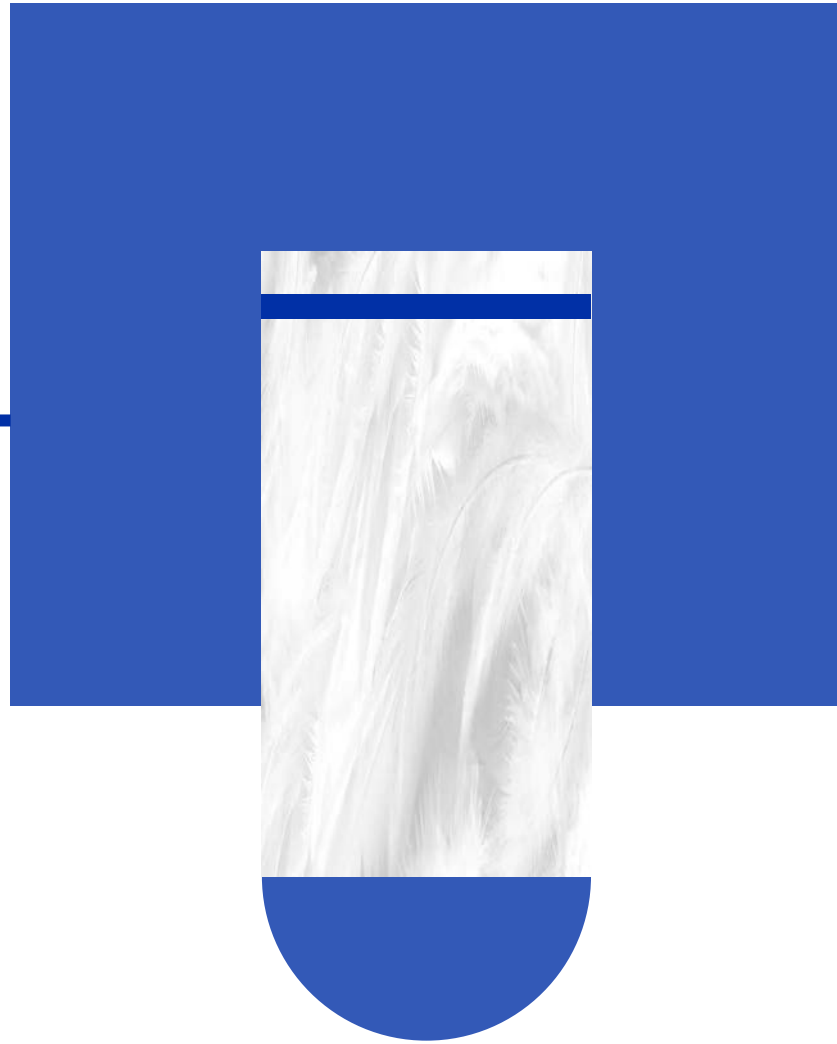
Summary

Statistical Measures of Asset Returns

Measures of shape of a distribution

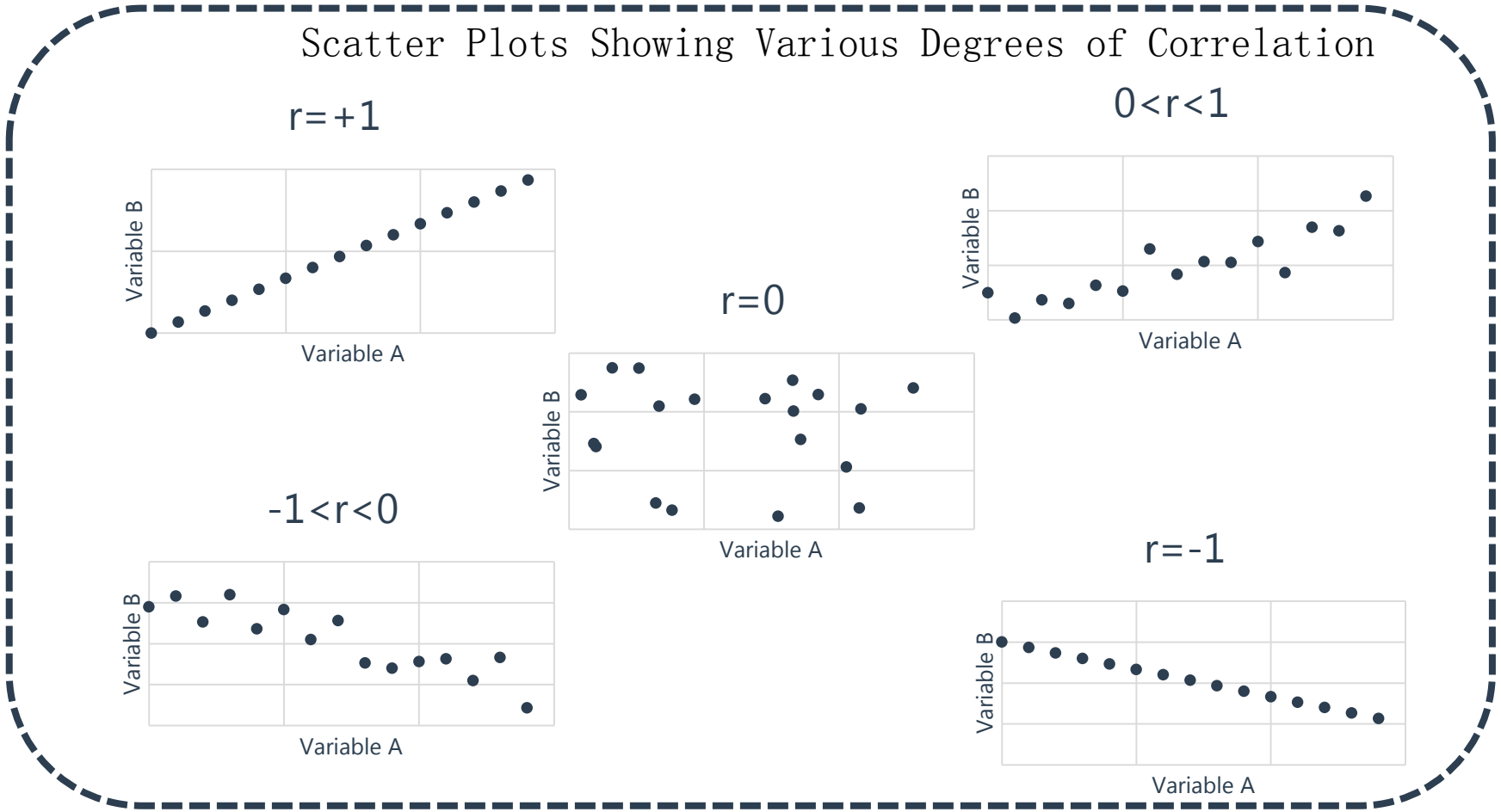
Correlation between two variables

- Interpret correlation between two variables to address an investment problem



Interpretations of Correlation Coefficients

- A **scatter plots** is a graph that shows the relationship between the observations for two data series in two dimensions.



Example

Properties of correlation

- Which of the following correlation coefficients indicates the weakest linear relationship between variables?

  A. -0.67.
  B. -0.24.
  C. 0.33.

- **Correct Answer: B.**
  - Correlation near +1 exhibit strong positive linearity, whereas correlations near -1 exhibit strong negative linearity. A correlation of 0 indicates an absence of any linear relationship between the variables. The closer the correlation is to 0, the weaker the linear relationship.

Limitations to Correlation Analysis

- Being cautious in basing investment strategies on high correlations, spurious correlations may suggest investment strategies that appear profitable but would not be, if implemented.
- **Three limitations of correlation analysis**
  - Nonlinear relationships
    - Two variables can have a strong nonlinear relation and still have a very low correlation, e.g., $Y = X^2$.
  - Outliers should be included because it contain information about the two variables' relationship. Otherwise, exclude the outliers.
  - Spurious correlation
    - correlation between two variables that reflects chance relationships in a particular dataset;
    - correlation induced by a calculation that mixes each of two variables with a third variable;
    - correlation between two variables arising not from a direct relation between them but from their relation to a third variable.

Summary

Statistical Measures of Asset Returns

Correlation between two variables

Summary

Module: Statistical Measures of Asset Returns

Measures of Central Tendency and Location

Measures of Dispersion

Measures of Shape of a Distribution

Correlation between Two Variables

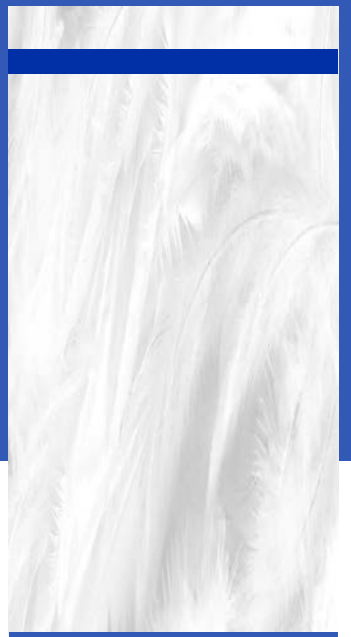
Module

Probability Trees and Conditional Expectations

1. Expected Value and Variance
2. Probability Trees and Conditional Expectations
3. Bayes' Formula and Updating Probability Estimates

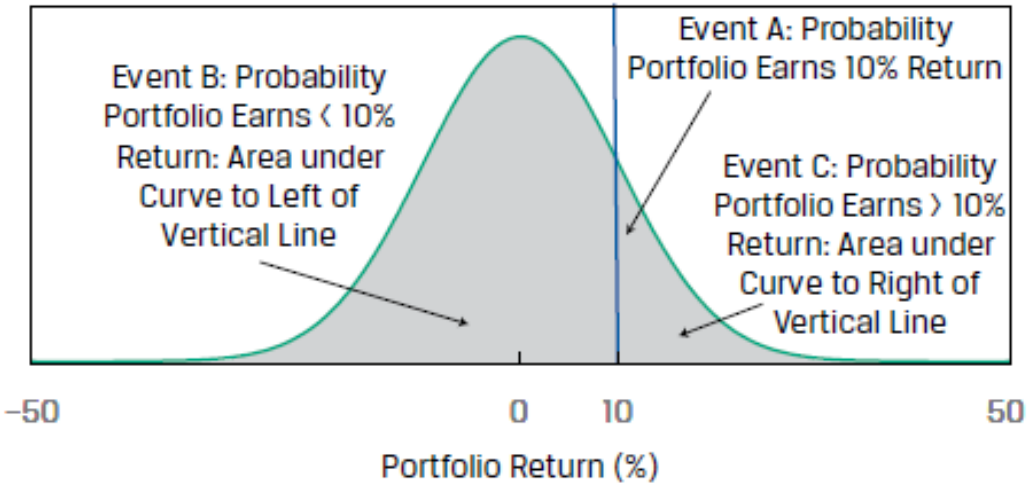
Expected value and variance

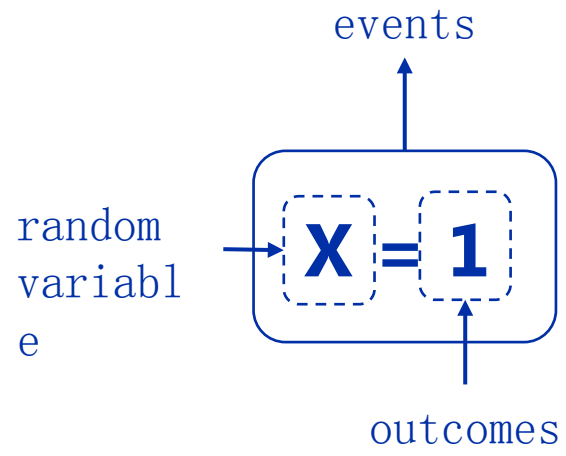
- Calculate expected values, variances, and standard deviations and demonstrate their application to investment problems



Probability Concepts

- **Basic Concepts**
  - A **random variable** is a quantity whose future outcomes are uncertain.
  - **Outcomes** are the possible values of a random variable.
  - An **event** is a specified set of outcomes.
- **Probability**
  - a number between 0 and 1 that measures the chance that a stated event will occur.

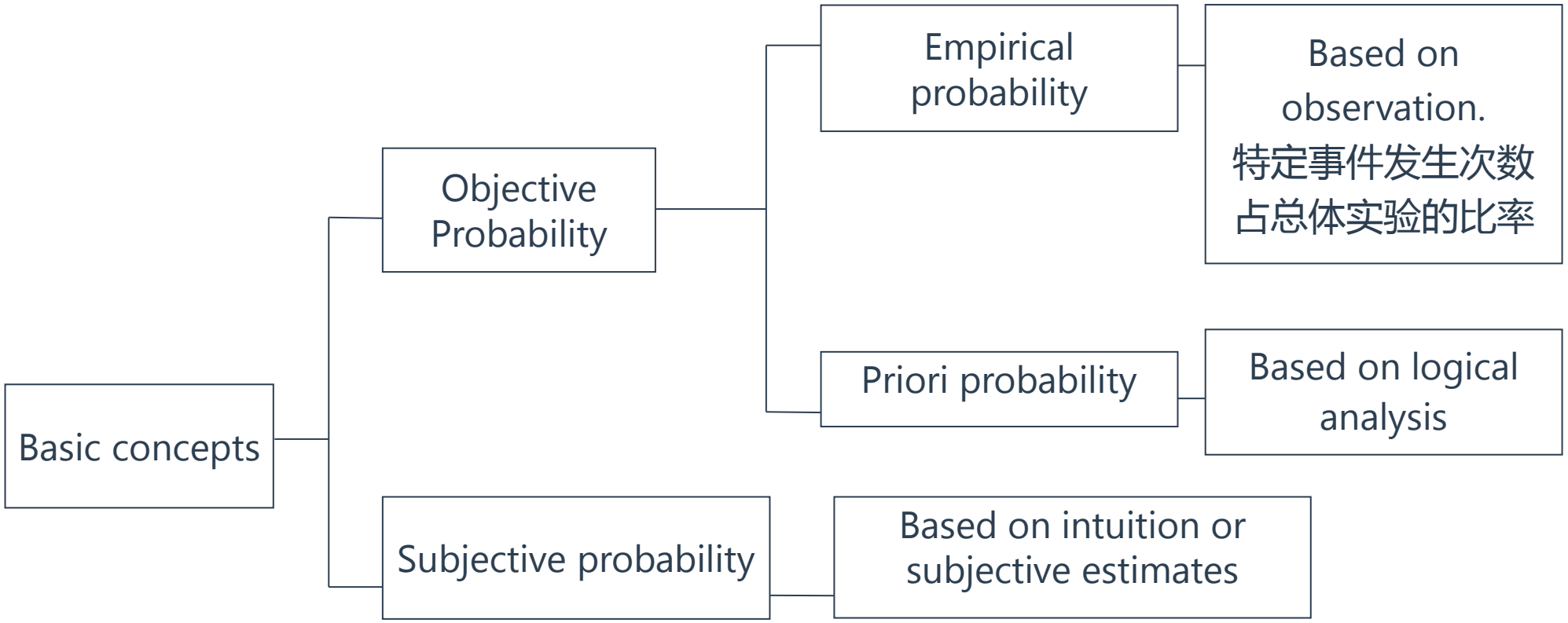




Probability Concepts

- Three approaches to estimate probabilities (e.g., What is the probability of flipping a coin and getting exactly the head out of two possible outcomes?)
  - **Empirical probability** (经验概率) is based on **observation**.
    - e.g., Analysts do the experiment 100 times (one flip each time) and find that the head get 46 times. The empirical probability is 46/100= 46%.
  - **Priori probability** (先验概率) is based on **logical analysis**.
    - e.g., Analysts assume the theoretical probability applies and the mathematical probability of the head out of two possible outcomes is 0.5 (1/2).
  - **Subjective probability** (主观概率) is based on **personal judgment**.
    - e.g., Analysts assume the probability is somewhere between 0.5 and 0.6, so analysts split the difference and choose 0.55.

Probability Concepts

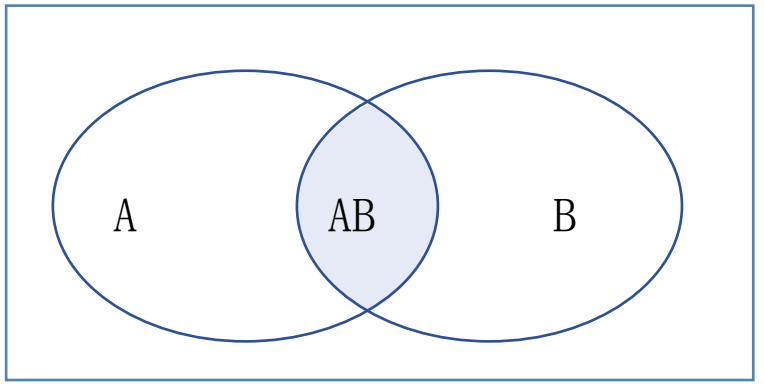


Probability Concepts

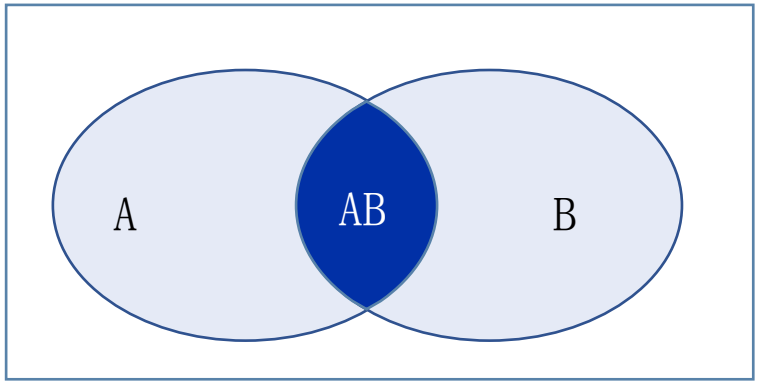
- **Unconditional Probability (marginal probability): P(A)**
  - What's the probability of event A?
- **Conditional probability: P(A|B)**
  - What's the probability of event A, given that B has occurred?

Calculation Rules for Probabilities

- **Multiplication rule**
  - Probability that two events will happen at the same time: Joint probability P(AB)
    - $P(AB)=P(A|B)\times P(B)= P(B|A)\times P(A)$



- **Addition rule**
  - The probability that A or B occurs, or both occur
    - P(A or B)=P(A)+P(B)-P(AB)



Calculation Rules for Probabilities

- **Mutually exclusive events**
  - P(AB)=P(A|B)=P(B|A)=0
  - P(A or B)=P(A)+P(B)
- **Independent events**
  - The occurrence of A doesn't affect the occurrence of B.
    - P(A|B)=P(A) or P(B|A)=P(B)
    - P(AB)=P(A)×P(B)
    - P(A or B)=P(A)+P(B)- P(AB)
- **Dependent events**
  - The probability of occurrence of A is related to the occurrence of B.

If exclusive, must NOT independence.

Expected Value & Variance

- **Expected value and variance of a random variable**
  - $\mathrm{E(X)} = \sum x_I{*}P(x_i) = x_1 * P(x_1) + x_2 * P(x_2) + \cdots + x_n * P(x_n)$
  - $\sigma^2 = \sum_{i=1}^{N} E\big(X - E(X)\big)^2$
  - $\sigma = \sqrt{\sigma^2}$

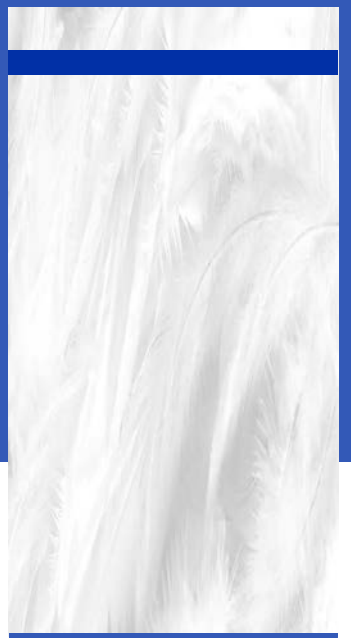
Summary

Probability Trees and Conditional Expectations

Expected Value and Variance

Probability trees and conditional expectations

- Formulate an investment problem as a probability tree and explain the use of conditional expectations in investment application



Total probability formula

- **Conditional expected values**
  - The expected value of a random variable X given an event or scenario S is denoted E(X | S)

$$E(X|S) = P(X_1|S)X_1 + P(X_2|S)X_2 + \ldots + P(X_n|S)X_n$$

- **Total probability formula for expected value**

$$E(X) = E(X|S_1)P(S_1) + E(X|S_2)P(S_2) + \ldots + E(X|S)P(S_n)$$

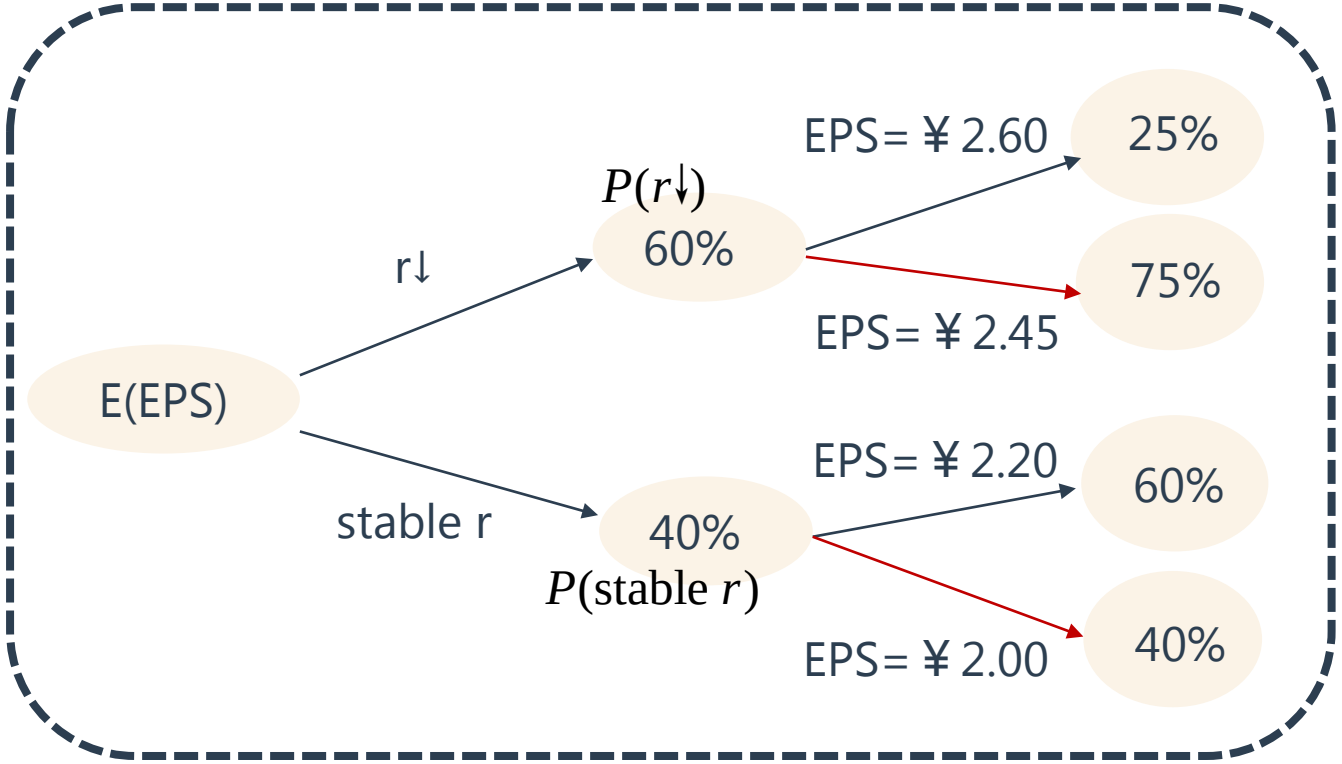
  - Where the set of events $\{S_1, S_2 \ldots S_n\}$ is mutually exclusive and exhaustive.

- **Conditional variance**

$$\sigma^2 = \sum_{i=1}^{N} P_i(X_i - E(X))^2$$

Example

Probability tree diagram



$P(\text{EPS}=2.60|\text{r}\downarrow)$

$P(\text{EPS}=2.45|\text{r}\downarrow)$

$P(EPS=2.45)=P(EPS=2.45|\text{r}\downarrow)\times P(r\downarrow)$

$P(\text{EPS}=2.20|\text{stable r})$

$P(\text{EPS}=2.00|\text{stable r})$

$$E(\text{EPS}|\text{r}\downarrow)=25\%\times 2.6+75\%\times 2.45=2.4875$$

$$E(\text{EPS}|\text{stable r})=60\%\times 2.20+40\%\times 2.00=2.12$$

$$E(\text{EPS})=2.4875\times 60\%+2.12\times 40\%=2.3405$$

$$\sigma^2(\text{EPS}|\text{r}\downarrow)$$

$$=25\%(2.6-2.4875)^2+75\%(2.45-2.4875)^2$$

$$=0.004219$$

$$\sigma^2(\text{EPS}|\text{stabler r})=0.0096$$

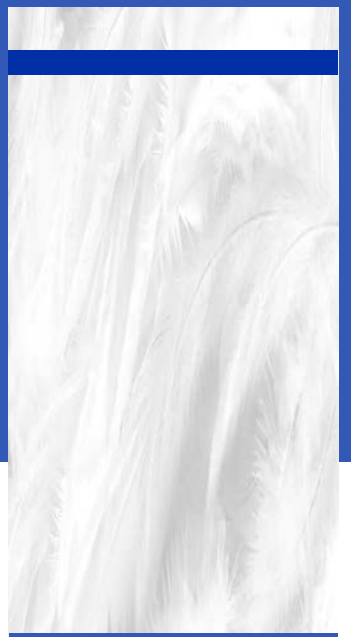
Summary

Probability Trees and Conditional Expectations

Probability Trees and Conditional Expectations

Bayes' formula and updating probability estimates

- Calculate and interpret an updated probability in an investment setting using Bayes' formula



Bayes' Formula

- $P(AB)=P(A|B)\times P(B) = P(B|A)\times P(A)$
- $\boldsymbol{P(A|B) = \frac{P(B|A)}{P(B)} \times P(A)}$
  - Where P(B) can be solved using **total probability formula**:
    - $P(B)=P(B|W_1)\times P(W_1)+P(B|W_2)\times P(W_2)+\ldots+ P(B|W_n)\times P(W_n)$
    - Wi is a set of mutually exclusive and exhaustive events.
- **Updated probability (posterior probability)** of event given the new information=

$$\frac{\text{Probability of the new information given event}}{\text{Unconditional probability of the new information}} \times \textbf{Prior probability} \text{ of event}$$

  - Equal prior probabilities are called **diffuse priors.**

Example

Bayes' Formula

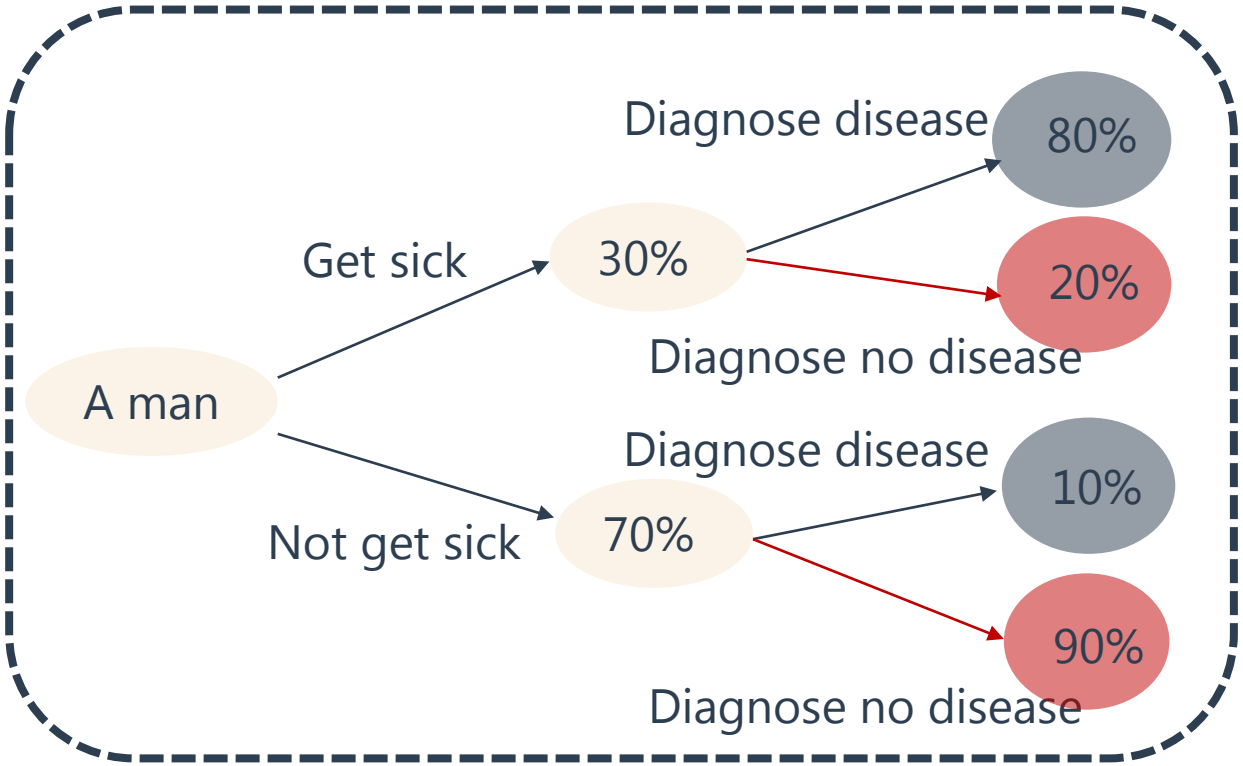
- The probability that a man has got sick is 30%, and if it does, a medical machine will have 80% chance to diagnose the disease. The probability that a man is healthy is 70%, and if it does, a medical machine will have 10% chance to diagnose the disease, and 90% chance to diagnose no disease. What's the probability that the man has actually got sick when the machine diagnoses disease?

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

$$= \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|\bar{A}) \times P(\bar{A})}$$

$$= \frac{30\% \times 80\%}{30\% \times 80\% + 70\% \times 10\%}$$

$$= 77.42\%$$



Example

- An analyst is screening a set of 100 stocks based on two criteria (Criterion 1 and Criterion 2). The passing level is that 50% of the stocks passed each screen. For these stocks, the values for Criterion 1 and Criterion 2 are not independent but are positively related. How many stocks should pass the analyst's screens?

  A. 25.
  B. More than 25.
  C. Less than 25.

- **Correct Answer: B.**
  - If the two criteria are independent, the joint probability of passing both screens 0.50×0.50= 0.25. However, the two criteria are positively related, then the contingent probability of P(pass Criterion 1 | pass Criterion B) is greater than 0.50. The joint probability of passing both screens is greater than 0.25.

Summary

Probability Trees and Conditional Expectations

Bayes' Formula and Updating Probability Estimates

Summary

Module: Probability Trees and Conditional Expectations

Expected Value and Variance

Probability Trees and Conditional Expectations

Bayes' Formula and Updating Probability Estimates

Module

Portfolio Mathematics

1. Portfolio Expected Return and Variance of Return
2. Forecasting Correlation of Returns: Covariance Given a Joint Probability
3. Portfolio Risk Measures: Applications of the Normal Distribution

Portfolio Expected Return and Variance of Return

- Calculate and interpret the expected value, variance, standard deviation, covariances, and correlations of portfolio returns

Expected return and variance of portfolios

- **Expected return, variance and standard deviation of a portfolio**
  - The expected return on the portfolio ($E(R_p)$) is a weighted average of the expected returns (R_1 to R_n) on the component securities using their respective proportions of the portfolio in currency units as weights (w_1 to w_n).

$$E(R_p) = \sum_{i=1}^{n} w_i E(R_i) = E(w_1 R_1 + w_2 R_2 + w_3 R_3 + \ldots + w_n R_n)$$

  - The portfolio variance of return is a measure of investment risk in a forward-looking sense.

$$\sigma^2(R_P) = E\left\{\left[R_p - E(R_p)\right]^2\right\} \qquad \sigma^2{}_p = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \operatorname{cov}(R_i, R_j)$$

Probability Concepts

- **Covariance**
  - Covariance is a measure of the co-movement between random variables.

$$\mathrm{COV}(X,Y) = \mathrm{E}[(X\text{-}\mathrm{E}(X))(Y\text{-}\mathrm{E}(Y))] \qquad Cov(R_i, R_j) = E\left[\left(R_i - ER_i\right)\left(R_j - ER_j\right)\right]$$

  - The covariance of a random variable with itself is its own variance.

$$\mathrm{COV}(X,X) = \mathrm{E}[(X - \mathrm{E}(X))(X - \mathrm{E}(X))] = \sigma^2(X)$$

  - Covariance ranges from negative infinity to positive infinity.
- **Correlation**
  - Correlation measures the co-movement (linear association) between two random variables.

$$\rho_{XY} = \frac{\mathrm{COV}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

  - Correlation is a number between −1 and +1.
  - Understand the difference between correlation and independence.
    - If ρ=0, there is **no linear relationship** between two variables.

Example

- An individual wants to invest $100,000 and is considering the following stocks:

stock	Expected Return	Standard Deviation of Returns
A	12%	15%
B	16%	24%

The expected correlation of returns for the two stocks is +0.5. If the investor invests $40,000 in Stock A and $60,000 in Stock B, the expected standard deviation of returns on the portfolio will be:

A. equal to 20.4%.

B. less than 20.4%.

C. greater than 20.4% because the correlation coefficient is greater than zero.

- **Correct Answer: B.**

Example

- An fund manager has a portfolio of two mutual funds, A and B, 75 percent invested in A, as shown in the following table.

Covariance Matrix		
Fund	A	B
A	625	120
B	120	196

The correlation between A and B, and the portfolio standard deviation of return is closest to:

	Correlation between A and B	Portfolio standard deviation of return
A.	0.18	40.80
B.	0.34	20.22
C.	0.12	18.00

- **Correct Answer: B.**

Summary

Portfolio Mathematics

Portfolio Expected Return and Variance of Return

Forecasting Correlation of Returns: Covariance Given a Joint Probability

- Calculate and interpret the covariance and correlation of portfolio returns using a joint probability function for returns

Example

Joint Probability Function

- The joint probability of the returns of Asset A and Asset B are given in the following figure.

Joint Probability Table			
Joint Probabilities	R_B=0.40	R_B=0.20	R_B=0.00
R_A=0.20	0.15	0	0
R_A=0.15	0	0.60	0
R_A=0.04	0	0	0.25

The covariance of returns for Asset A and Asset B is closest to:

A. 0.0003.

B. 0.0024.

C. 0.0066.

Example

Joint Probability Function

- **Correct Answer: C.**
  - The expected returns for the individual assets are determined as:

$$E(R_A)=P(R_{A1},R_{B1})R_{A1}+P(R_{A2},R_{B2})R_{A2}+P(R_{A3},R_{B3})R_{A3}$$

$$E(R_A)=(0.15)(0.20)+(0.60)(0.15)+(0.25)(0.04)=0.13$$

$$E(R_B)=P(R_{B1},R_{A1})R_{B1}+P(R_{B2},R_{A2})R_{B2}+P(R_{B3},R_{A3})R_{B3}$$

$$E(R_B)=(0.15)(0.40)+(0.60)(0.20)+(0.25)(0.00)=0.18$$

  - The covariance of the asset returns is determined as:

$$Cov(R_A,R_B)$$

$$=P(R_{A1},R_{B1})[(R_{A1}-E(R_A)]\ [(R_{B1}-E(R_B)]$$

$$+P(R_{A2},R_{B2})[(R_{A2}-E(R_A)]\ [(R_{B2}-E(R_B)]$$

$$+P(R_{A3},R_{B3})[(R_{A3}-E(R_A)]\ [(R_{B3}-E(R_B)]$$

$$=0.15(0.20-0.13)(0.40-0.18)+0.6(0.15-0.13)(0.20-0.18)$$

$$+0.25(0.04-0.13)(0.00-0.18)=0.0066$$

Summary

Portfolio Mathematics

Forecasting Correlation of Returns: Covariance Given a Joint Probability

Portfolio Risk Measures: Applications of the Normal Distribution

- Define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy's safety-first criterion

*Basic Concepts

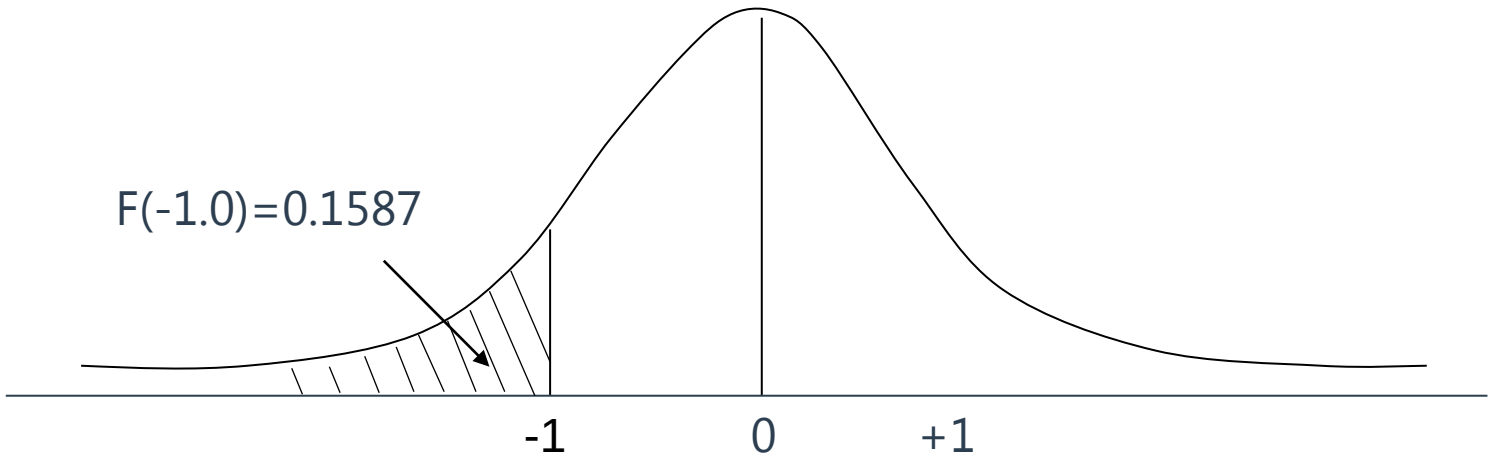
- **Probability Distribution**
  - Specifies the probabilities of the possible outcomes of a random variable.
- **Discrete and continuous random variables**
  - **Discrete random variables** take on at most a **countable** number of possible outcomes **but do not necessarily to be limited**.
  - **Continuous random variables**: cannot describe the possible outcomes of a continuous random variable Z with a list z_1, z_2, . . . because the outcome $(z_1 + z_2)/2$, not in the list, would always be possible.
    - P (x)=0 even though x can happen.
    - P $(x_1<X<x_2)$

*Basic Concepts

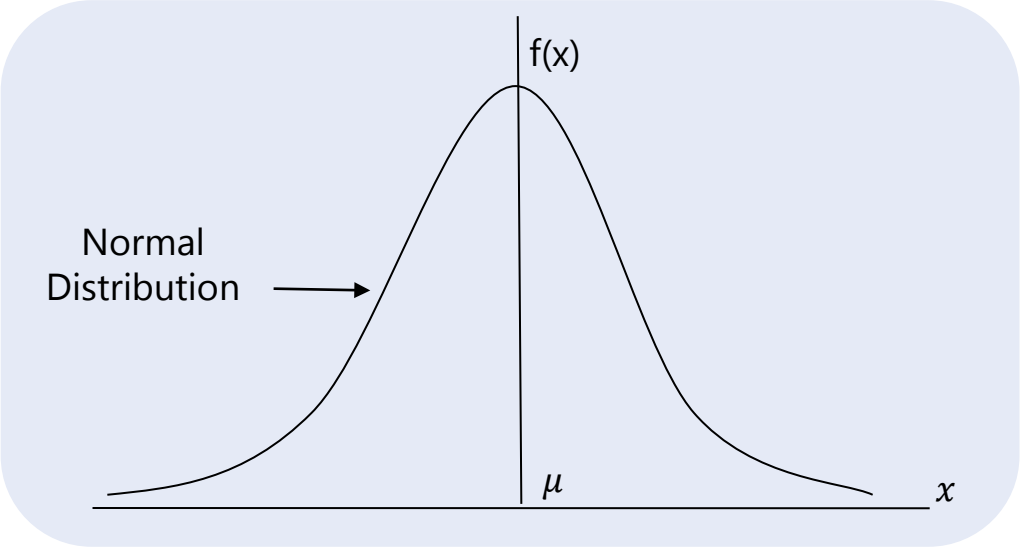
- **Probability function:** $p(x)=P(X=x)$
  - For discrete random variables
  - $0 \leq p(x) \leq 1$
  - $\Sigma p(x)=1$
- **Probability density function (p.d.f) : f(x)**
  - For continuous random variable commonly
- **Cumulative probability function (c.p.f) : F(x)**
  - $F(x)=P(X<=x)$

*Basic Concepts

- **Probability density function**



*Normal Distribution



- **X~N $(\mu,\ \sigma^2)$**
- **Symmetrical distribution: Skewness=0; kurtosis=3**
- **A linear combination of normally distributed random variables is also normally distributed.**
- **As the values of x gets farther from the mean, the probability density get samller and smaller but are always positive.**

*Normal distribution

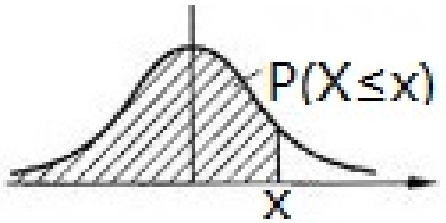
- **Standard normal distribution**
  - N(0,1) or Z
  - Standardization: if X~N (μ, σ^2), then Z=$\frac{X-\mu}{\sigma}$ ~ N(0,1)
  - Z-table
    - F(-z)=1-F(z)
    - P(Z>z)=1-F(z)

*Z-table

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7582
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767

Cumulative Probabilities for a Standard Normal Distribution

$P(X \leq x) = F(x)$ for $x \geq 0$



Mean–variance analysis

- In economic theory, **mean–variance analysis** holds exactly when investors are risk averse; when they choose investments to maximize expected utility or satisfaction; and when either (assumption 1) returns are normally distributed or (assumption 2) investors have quadratic utility functions (a concept used in economics for a mathematical representation of risk and return trade-offs).
- **Shortfall risk: R_L = threshold level return, minimum return required**
  - Minimize ($R_p < R_L$) $[E(R_P) - R_L]/\sigma_P$
- **Roy's safety-first criterion**
  - $$SFR = \frac{E(R_P) - R_L}{\sigma_P}$$
- **Maximize S-F-Ratio**
  - Maximize <=> Minimize P ($Rp < R_L$)

Managing financial risk tools

- **Stress testing and scenario analysis**
  - Refer to a set of techniques for estimating losses in extremely unfavorable combinations of events or scenarios.
  - Scenario analysis: A technique for exploring the performance and risk of investment strategies in different structural regimes.
  - Stress testing A specific type of scenario analysis that estimates losses in rare and extremely unfavorable combinations of events or scenarios.
- **Value at risk (VaR)**
  - A money measure of the minimum value of losses expected over a specified time period (e.g., a day, a quarter, or a year) at a given level of probability (often 0.05 or 0.01).

Example

- A portfolio manager gathered the following information about four possible asset allocations:

Allocation	Expected annual return	Standard deviation of return
A	10%	6%
B	25%	14%
C	18%	17%

The manager's client has stated that her minimum acceptable return is 8%. Based on Roy's safety-first criterion, the *most* appropriate allocation is:

A. Allocation A.
B. Allocation B.
C. Allocation C.

- **Correct Answer: B**

Example

- You are researching asset allocations for a client with an $1,000,000 portfolio. Although her investment objective is long-term growth, at the end of a year she may want to liquidate $40,000 of the portfolio to fund educational expenses. If that need arises, she would like to be able to take out he $40,000 without invading the initial capital of $1,000,000. The following table shows three alternative allocations.

	A	B	C
Expected annual return	26	13	15
Standard deviation of return	28	9	21

Address these questions (assume normality for Parts 2 and 3):

1. Given the client's desire not to invade the $1,000,000 principal, what is the shortfall level, R_L? Use this shortfall level to answer Part 2.
2. According to the safety-first criterion, which of the three allocations is the best?
3. What is the probability that the return on the safety-first optimal portfolio will be less than the shortfall level? (F(1.00)=0.8413)

Example

- Correct Answer:

1. R_L=40,000/1,000,000=4.00%
2. A: SFR_A=(26-4)/28=0.79

   B: SFR_B=(13-4)/9=1.00

   C: SFR_C=(15-4)/21=0.52

   Allocations B is best one because it has the highest SFR.

3. $P(R_B<4.00)=P[(R_B-13)/9 < (4.00-13)/9] = F(-1.00)$

   $= 1-F(1.00)=1-0.8413=0.1587\approx16\%$

The safety-first optimal portfolio has a roughly 16% chance of not meeting a 4.00% return threshold.

Summary

Portfolio Mathematics

Portfolio Risk Measures: Applications of the Normal Distribution

Summary

Module: Portfolio Mathematics

Portfolio Expected Return and Variance of Return

Forecasting Correlation of Returns: Covariance Given a Joint Probability

Portfolio Risk Measures: Applications of the Normal Distribution

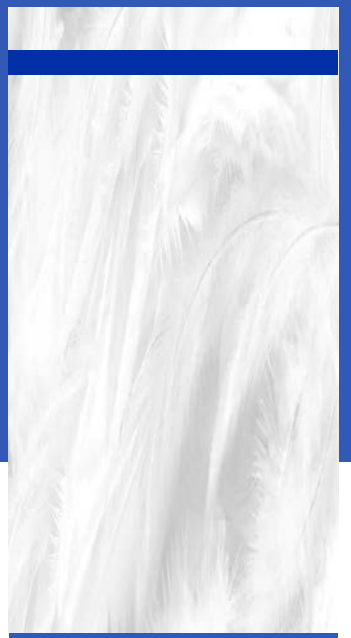
Module

Simulation Methods

1. Lognormal distribution and continuous compounding
2. Monte Carlo Simulation

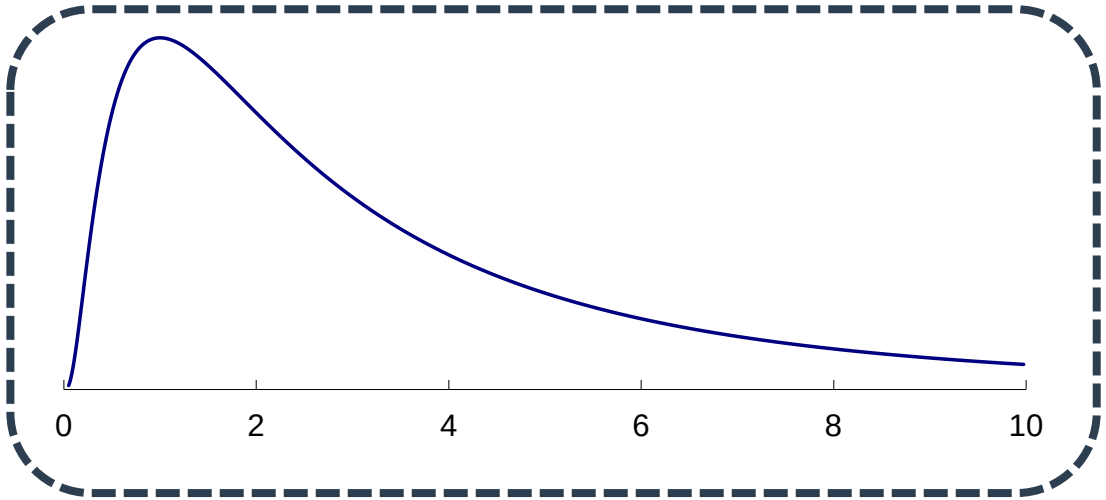
Lognormal distribution and continuous compounding

- 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



Lognormal Distribution

- **Lognormal Distribution:**
  - If lnX is normal, then X is lognormal, which is used to describe the price of asset.
- **Features**
  - Right skewed.
  - The values of random variables that follow lognormal distribution are always be positive, so it is useful for modeling asset prices.



Continuously Compounded Rates of Return

- **A key assumption**
  - in many investment applications is that returns are independently and identically distributed (i.i.d.)
- Relationship between the distribution of **stock return and stock price**

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

$$r_{0,T} = r_{T-1,T} + r_{T-2,T} + \ldots + r_{0,1}$$

$$\longrightarrow$$

$$E\left(r_{0,T}\right) = E\left(r_{T-1,T}\right) + E\left(r_{T-2,T}\right) + \ldots + E\left(r_{0,1}\right) = \mu \times T$$

$$\sigma^2\left(r_{0,T}\right) = \sigma^2 T \text{ or } \sigma\left(r_{0,T}\right) = \sigma\sqrt{T}$$

by convention, it is stated as an annualized measure

- **Key assumption**
  - returns are independently and identically distributed (i.i.d.).
  - **Independence** captures the proposition that investors cannot predict future returns using past returns. **Identical distribution** captures the assumption of stationarity, a property implying that the mean and variance of return do not change from period to period.

Example

- Compared to a normal distribution, a lognormal distribution is *least likely* to be:

  A. Skewed to the left.
  B. Skewed to the right.
  C. Useful in describing the distribution of stock prices.

- **Solution: A**

- An analyst stated that lognormal distribution are suitable for describing asset returns and that normal distributions are suitable for describing distributions of asset prices. Is the analyst's statement correct with respect to:

	Lognormal distribution	Normal distribution
A.	No	No
B.	No	Yes
C.	Yes	No

- **Solution: A**

Example

Volatility of Share Price

- Suppose you are researching Astra International (Indonesia Stock Exchange: ASII) and are interested in Astra's price action in a week in which international economic news had significantly affected the Indonesian stock market. You decide to use volatility as a measure of the variability of Astra shares during that week. The following Exhibit shows closing prices during that week.

Astra International Daily Closing Prices

Day	Closing Price (Indonesian rupiah, IDR)
Monday	6,950
Tuesday	7,000
Wednesday	6,850
Thursday	6,600
Friday	6,350

- Use the data provided to do the following:

1. Estimate the volatility of Astra shares. (Annualize volatility on the basis of 250 trading days in a year.)

Example

Volatility of Share Price

- **Solution to Q1:**
  - First, calculate the continuously compounded daily returns; then, find their standard deviation in the usual way. In calculating sample variance, to get sample standard deviation, the divisor is sample size minus 1.
  - ln(7,000/6,950) = 0.007168; ln(6,850/7,000) = −0.021661; ln(6,600/6,850) = −0.037179; ln(6,350/6,600) = −0.038615.
  - Sum = −0.090287; Mean = −0.022572; Variance = 0.000452; Standard deviation = 0.021261.
  - The standard deviation of continuously compounded daily returns is 0.021261. In this example, σ is the sample standard deviation of one-period continuously compounded returns. Thus, σ refers to 0.021261. We want to annualize, so the horizon T corresponds to one year. Because σ is in days, we set T equal to the number of trading days in a year (250). Therefore, we find that annualized volatility for Astra stock that week was 33.6 percent, calculated as $0.021261\sqrt{250} = 0.336165$.

Volatility of Share Price

2. Calculate an estimate of the expected continuously compounded annual return for Astra.

- **Solution to Q2:**
  - Note that the sample mean, −0.022572 (from the Solution to Q1), is a sample estimate of the mean, μ, of the continuously compounded one-period or daily returns. The sample mean can be translated into an estimate of the expected continuously compounded annual return, $\mu \times T = -0.022572 (250)$ (using 250 to be consistent with the calculation of volatility).

Volatility of Share Price

3. Discuss why it may not be prudent to use the sample mean daily return to estimate the expected continuously compounded annual return for Astra.

- **Solution to Q3:**
  - Four daily return observations are far too few to estimate expected returns. Further, the variability in the daily returns overwhelms any information about expected return in a series this short.

4. Identify the probability distribution for Astra share prices if continuously compounded daily returns follow the normal distribution.

- **Solution to Q4:**
  - Astra share prices should follow the lognormal distribution if the continuously compounded daily returns on Astra shares follow the normal distribution.

Summary

Simulation Methods

Lognormal distribution and continuous compounding

Monte Carlo simulation

- Describe Monte Carlo simulation and explain how it can be used in investment applications

Monte Carlo simulation

- **Monte Carlo simulation** is the generation of a very large number of random samples from a specified probability distribution or distributions to obtain the likelihood of a range of results.
- **Limitations:**
  - The operating of Monte Carlo simulation is very complex and we must assume a parameter distribution in advance.
  - Monte Carlo simulation provides only statistical estimates, not exact results.

Example

Monte Carlo simulation

- Define Monte Carlo simulation and explain its use in investment management.

- Solution:
  - A Monte Carlo simulation generates of a large number of random samples from a specified probability distribution (or distributions) to represent the role of risk in the system.
  - Monte Carlo simulation is widely used to estimate risk and return in investment applications. In this setting, we simulate the portfolio's profit and loss performance for a specified time horizon.
  - Repeated trials within the simulation produce a simulated frequency distribution of portfolio returns from which performance and risk measures are derived.
  - Another important use of Monte Carlo simulation in investments is as a tool for valuing complex securities for which no analytic pricing formula is available. It is also an important modeling resource for securities with complex embedded options.

Example

Monte Carlo simulation

- Compared with analytical methods, what are the strengths and weaknesses of using Monte Carlo simulation for valuing securities?

- Solution:
  - Strengths: Monte Carlo simulation can be used to price complex securities for which no analytic expression is available, particularly European-style options.
  - Weaknesses: Monte Carlo simulation provides only statistical estimates, not exact results. Analytic methods, when available, provide more insight into cause-and-effect relationships than does Monte Carlo simulation.

Summary

Simulation Methods

Monte Carlo simulation

Summary

Module: Simulation Methods

Lognormal distribution and continuous compounding

Monte Carlo simulation

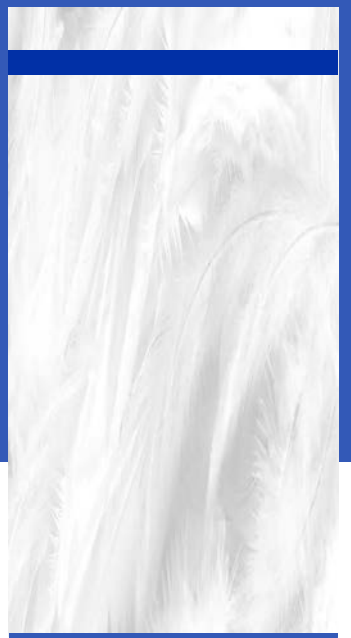
Module

Sampling and Estimation

1. Sampling methods
2. Central limit theorem and inference
3. Bootstrapping and Empirical Sampling Distributions

Sampling methods

- Compare and contrast simple random, stratified random, cluster, convenience, and judgmental sampling and their implications for sampling error in an investment problem



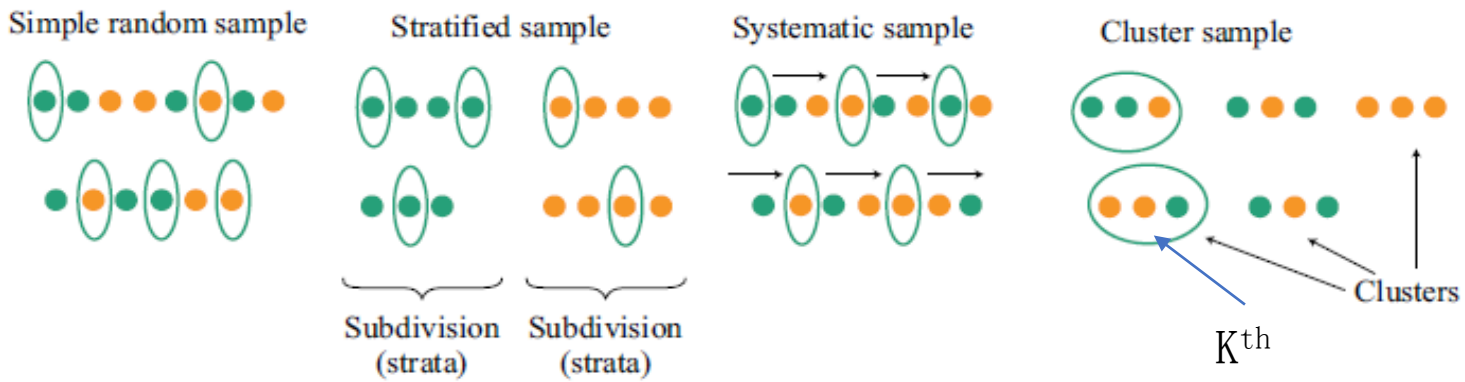
Statistical Concepts

- **Population**
  - A population is defined as all members of a specified group.
  - A **parameter** is used to describe the features of a population.
- **Sample**
  - A sample is a subset of a population.
  - A **sample statistic** is used to describes the features of a sample.

	Abbreviation (Parameter)	Abbreviation (Sample) Statistic
Mean	μ	$\bar{X}$
Variance	σ^2	s^2
Standard Deviation	σ	s

Sampling Methods

- Sampling Methods
  - Probability Methods
    - Simple Random Sampling
    - Systematic Sampling
    - Stratified Random Sampling
    - Cluster Sampling
  - Non-Probability Methods
    - Convenience Sampling
    - Judgment Sampling



Sampling Methods

- **Probability sampling** gives every member of the population an equal chance of being selected.
  - **A sampling plan** is the set of rules used to select a sample.
  - **Simple random sample** has each element of the population has an equal probability of being selected to the subset.
  - **Systematic sampling** selects every k^{th} member with a desired sample size.
  - **Stratifed Random Sampling (E.g., bond indexing)**
    - With classification criteria, dividing population into subpopulations/strata;
    - Drawing simple random samples from each stratum in sizes proportional to the relative size of each stratum in the population.
  - **Cluster Sampling**
    - Dividing population into clusters, a mini-representation of the entire populations;
    - Drawing simple random sampling to get certain clusters.

Sampling Methods

- **Non-probability sampling**: depends on factors (such as a sampler's judgment or the convenience) other than probability considerations to access data.
  - **Convenience Sampling**
    - an element is selected based on whether or not it is accessible to a researcher or on how easy it is for a researcher to access the element.
  - **Judgmental sampling**
    - selectively handpicking elements based on a researcher's knowledge and professional judgment, e.g., auditor focus on specific items in financial statement and related files.
    - Sample selection under judgmental sampling can be affected by the bias of the researcher and might lead to skewed results that do not represent the whole population.

Basic Concept

- **Sampling error:** difference between the observed value of a statistic and the quantity it is intended to estimate as a result of using subsets of the population.
  - E.g., Sampling error of the mean= sample mean - population mean
- The **sample statistic** itself is a random variable and has a probability distribution.
  - **Sampling distribution** of a statistic is the distribution of all the distinct possible values that the statistic can assume when computed from samples of the same size randomly drawn from the same population.

Example
Calculating Sharpe Ratios: One or Two Years of Quarterly Data

- Analysts often use the Sharpe ratio to evaluate the performance of a managed portfolio.
- The Sharpe ratio is the average return in excess of the risk-free rate divided by the standard deviation of returns. This ratio measures the return of a fund or a security above the risk-free rate (the excess return) earned per unit of standard deviation of return.
- To compute the Sharpe ratio, suppose that an analyst collects eight quarterly excess returns (i.e., total return in excess of the risk-free rate). During the first year, the investment manager of the portfolio followed a low-risk strategy, and during the second year, the manager followed a high-risk strategy. For each of these years, the analyst also tracks the quarterly excess returns of some benchmark against which the manager will be evaluated. For each of the two years, the Sharpe ratio for the benchmark is 0.21. The following exhibit gives the calculation of the Sharpe ratio of the portfolio.

Example
Calculating Sharpe Ratios: One or Two Years of Quarterly Data

Calculation of Sharpe Ratios: Low-Risk and High-Risk Strategies			
Quarter/Measure	Year 1 Excess Returns	Year 2 Excess Returns	Year 1~2 Excess Returns
Quarter 1	-3%	-12%	
Quarter 2	5%	20%	
Quarter 3	-3%	-12%	
Quarter 4	5%	20%	
Quarterly average	1%	4%	2.50%
Quarterly standard deviation	4.62%	18.48%	12.57%
Sharpe ratio	22%	22%	19.90%

- The second year's results (Sharpe ratio= 4/18.48 = 0.22) mirror the first year (SR=1/4.62%=0.22) except for the higher average return and volatility.
- During the first and second years, larger Sharpe ratios are better than smaller ones (providing more return per unit of risk), the manager appears to have outperformed the benchmark (SR=0.21).
- When returns for the two-year period are pooled, the manager appears to have provided less return per unit of risk than the benchmark and less when compared with the separate yearly results.

Example
Calculating Sharpe Ratios: One or Two Years of Quarterly Data

- The problem with using eight quarters of return data is that the analyst has violated the assumption that the sampled returns come from the same population.
- Combining the results for the first and second years yielded a sample that was representative of no population. Because the larger sample did not satisfy model assumptions, any conclusions the analyst reached based on the larger sample are incorrect. For this example, she was better off using a smaller sample than a larger sample because the smaller sample represented a more homogeneous distribution of returns.

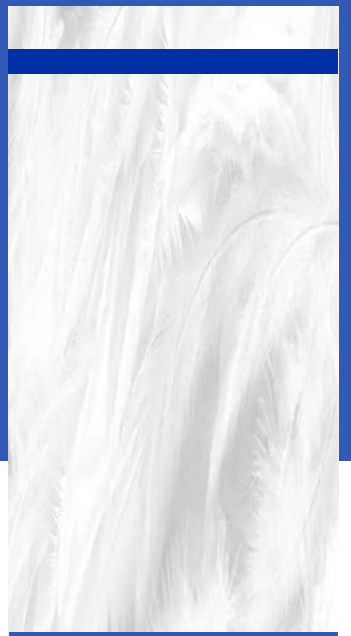
Summary

Sampling and Estimation

Sampling methods

Central Limit Theory

- Explain the central limit theorem and its importance for the distribution and standard error of the sample mean



Central Limit Theory

- **Central Limit Theory**
  - For simple random samples of size n from a population with a mean μ and a finite variance σ^2 but without known distribution, the sampling distribution of the sample mean approaches $N(\mu , \sigma^2/n)$ if the sample size is sufficiently large (generally n ≥ 30).
  - 条件：1. n≥ 30　　2. 总体均值、方差存在
  - 结论: 1.服从正态分布　2. $\mu_{\bar{X}} = \mu_X = \mu$　Var $(\bar{X})=\frac{\sigma_x^2}{n}$
- **Standard error of the sample mean**
  - Known population variance $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
  - Unknown population variance $s_{\bar{x}} = s/\sqrt{n}$

Summary

Sampling and Estimation

Central limit theorem and inference

Bootstrapping and empirical sampling distributions

- Describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

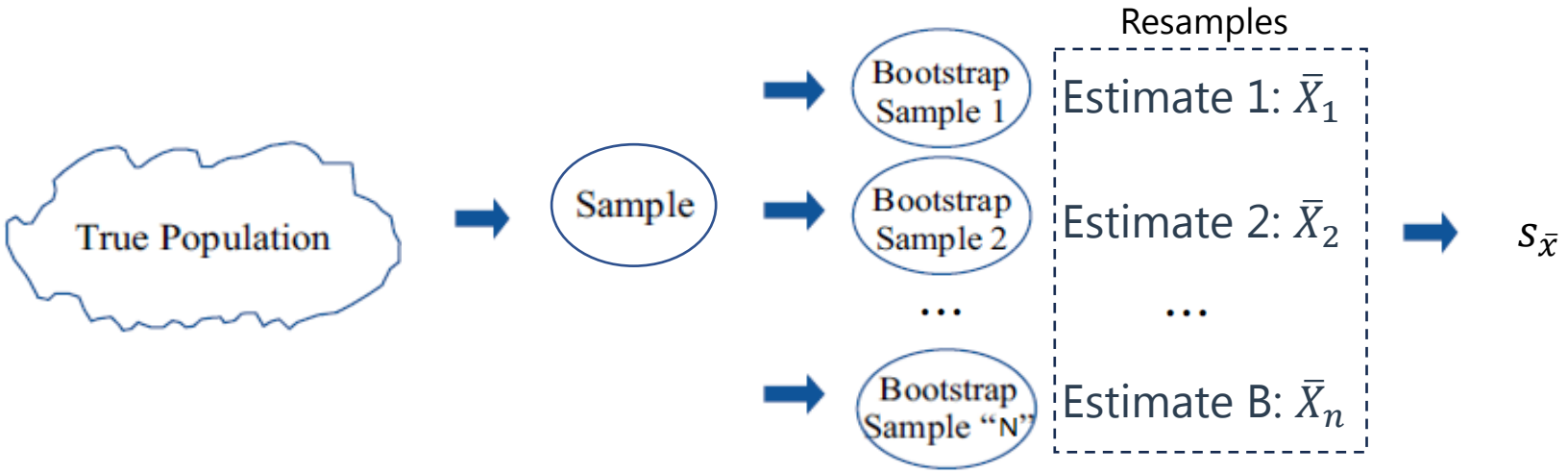
Resampling

- **Resampling:** repeatedly draws samples from the original observed data sample for the statistical inference of population parameters.
- **Two types of resampling**
  - **Bootstrapping** uses computer simulation for statistical inference without using an analytical formula such as a z-statistic or t-statistic (model-free resampling or non-parametric resampling)
  - While bootstrap repeatedly draws samples with replacement, **jackknife** samples are selected by taking the original observed data sample and leaving out one observation at a time from the set (and not replacing it).

Bootstrapping

- **Process**

1. Mimics the process by treating the randomly drawn sample as if it were the population;
2. **Repeatedly draw** samples from the original sample. (Each resample is of **the same size as the original sample**.)
3. Constructs resamples and calculate the estimates.
4. Calculate the standard error of the estimator: $s_{\bar{x}} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\bar{X} - E\bar{X})^2}$



Bootstrapping & Jackknife

- **Bootstrap**
  - to estimate the standard error of a sample mean
  - to find the standard error or construct confidence intervals for the statistic of other population parameters, such as the median
  - Compared with conventional statistical methods, bootstrap does not rely on an analytical formula to estimate the distribution of the estimators
  - bootstrap can be applied widely in finance, such as for historical simulations in asset allocation or in gauging an investment strategy’s performance against a benchmark
- **Jackknife**
  - reduce the bias of an estimator
  - find the standard error and confidence interval of an estimator

Jackknife vs. Bootstrap

- **Jackknife vs. Bootstrap**
  - **Jackknife** produces similar results for every run.
    - Jackknife usually requires n repetitions. (n=sample size).
  - **Bootstrap** usually gives different results because bootstrap resamples are randomly drawn.
    - Bootstrap needs to determine how many repetitions are appropriate.

Example

- An analyst in a real estate investment company is researching the housing market of the Greater Boston area. From a sample of collected house sale price data in the past year, she estimates the median house price of the area. To find the standard error of the estimated median, she is considering two options:

  Option 1 The standard error of the sample median can be given by $\frac{S}{\sqrt{n}}$, where s denotes the sample standard deviation and n denotes the sample size.

  Option 2 Apply the bootstrap method to construct the sampling distribution of the sample median, and then compute the standard error.

  Which of the following statements is accurate?

  A. Option 1 is suitable to find the standard error of the sample median.

  B. Option 2 is suitable to find the standard error of the sample median.

  C. Both options are suitable to find the standard error of the sample median.

Example

- **Correct Answer: B.**
  - Option 1 is valid for estimating the standard error of the sample mean but not for that of the sample median, which is not based on the given formula. Thus, both A and C are incorrect. The bootstrap method is a simple way to find the standard error of an estimator even if no analytical formula is available or it is too complicated.

Summary

Sampling and Estimation

Bootstrapping and Empirical Sampling Distributions

Summary

Module: Sampling and Estimation

Sampling methods

Central limit theorem and inference

Bootstrapping and Empirical Sampling Distributions

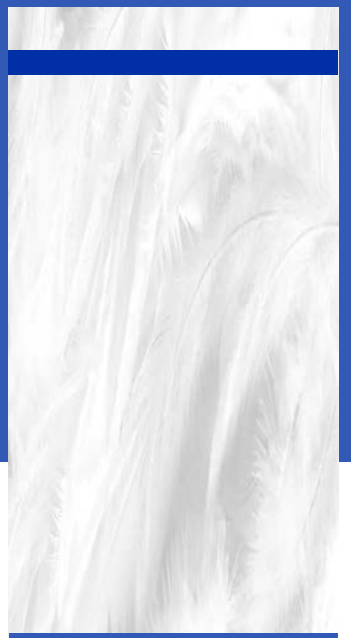
Module

Hypothesis Testing

1. Hypothesis tests for finance
2. Tests of return and risk in finance
3. Parametric versus Nonparametric Tests

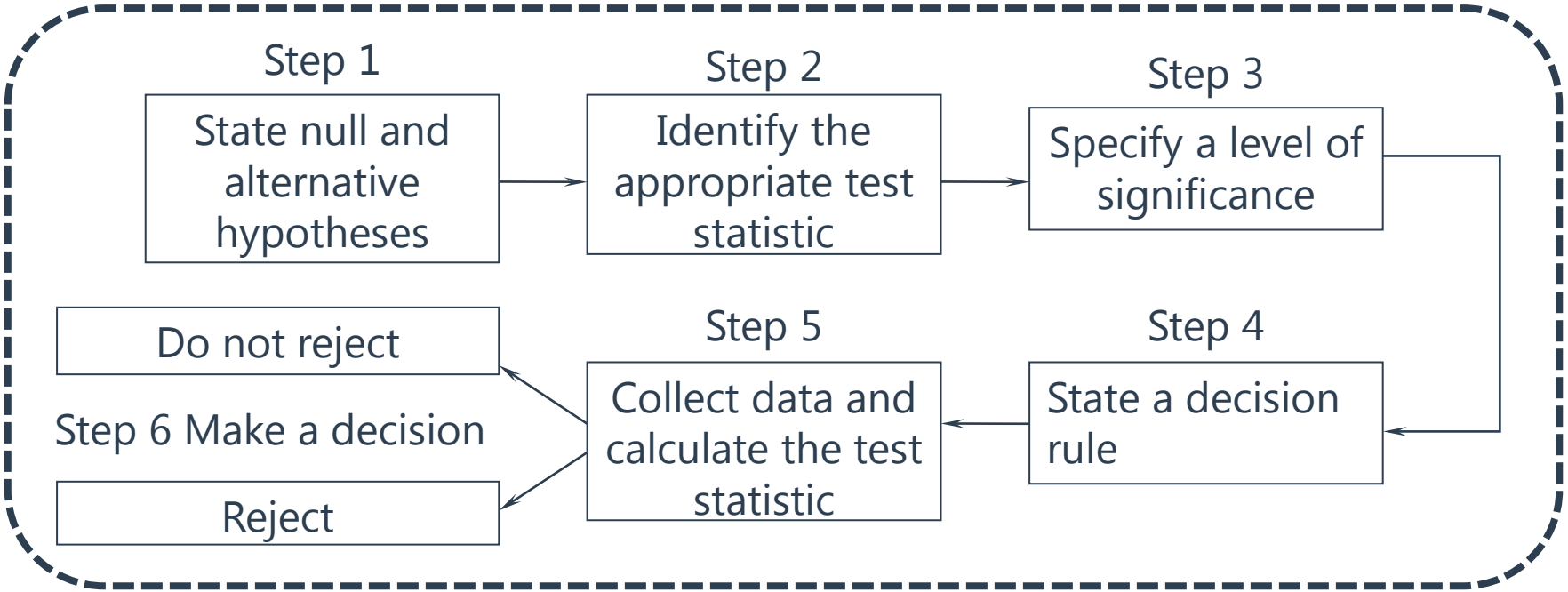
Hypothesis tests for finance

- Explain hypothesis testing and its components, including statistical significance, Type I and Type II errors, and the power of a test.



Hypothesis Testing

- **The process of hypothesis testing**



- Differences between **statistically significant** and **economically meaningful**.
  - Although a strategy may provide a statistically significant positive mean return, the results may not be economically significant when accounting for transaction costs, taxes, and risk.

Hypothesis Testing

- **Step one: state the hypothesis**
  - Statistical assessment of a statement or idea regarding a population parameter.
  - Null hypothesis and Alternative hypothesis (we want to assess)
    - The fact we suspect and want to reject
    - For population not sample

$$H_0 : \mu = \mu_0 \qquad H_a : \mu \neq \mu_0$$

Two-tailed $H_0 : \mu = \mu_0 \qquad H_a : \mu \neq \mu_0$

One-tailed

$$H_0 : \mu \leq \mu_0 \qquad H_a : \mu > \mu_0$$
$$or, H_0 : \mu \geq \mu_0 \qquad H_a : \mu < \mu_0$$

Example

- In the hypothesis testing, assess whether if mean excess the benchmark, how to set the null hypothesis?

A. $\mu < \mu_0$

B. $\mu \leq \mu_0$

C. $\mu > \mu_0$

- **Correct Answer: B**

Example

- Austin Roberts believes that the mean price of houses in the area is greater the $145,000. the appropriate alternative hypothesis is:

  A. H_a: $\mu < \$145,000$
  B. H_a: $\mu \geq \$145,000$
  C. H_a: $\mu > \$145,000$

- **Correct Answer: C.**

- An analyst is conducting a hypothesis test to determine if the mean time spent on investment research is different from three hours per day. The appropriate null hypothesis for the described test is:

  A. H_0: $\mu = 3$ hours, two-tailed test.
  B. H_0: $\mu = 3$ hours, one-tailed test.
  C. H_0: $\mu \geq 3$ hours, two-tailed test.

- **Correct Answer: A.**

Hypothesis Testing

- **Step two: identify the appropriate test statistic**
  - Test statistic $= \frac{Sample\ statistics - Hypothesized\ value}{stanard\ error\ of\ the\ sample\ statistic}$

  - Test Statistic follows Normal, t, Chi Square or F-distributions
  - Test Statistic has formula. Calculate it with the sample data. We should emphasize Test Statistic is calculated by ourselves not from the table.
  - This is the general formula but only for Z and t-distributions.

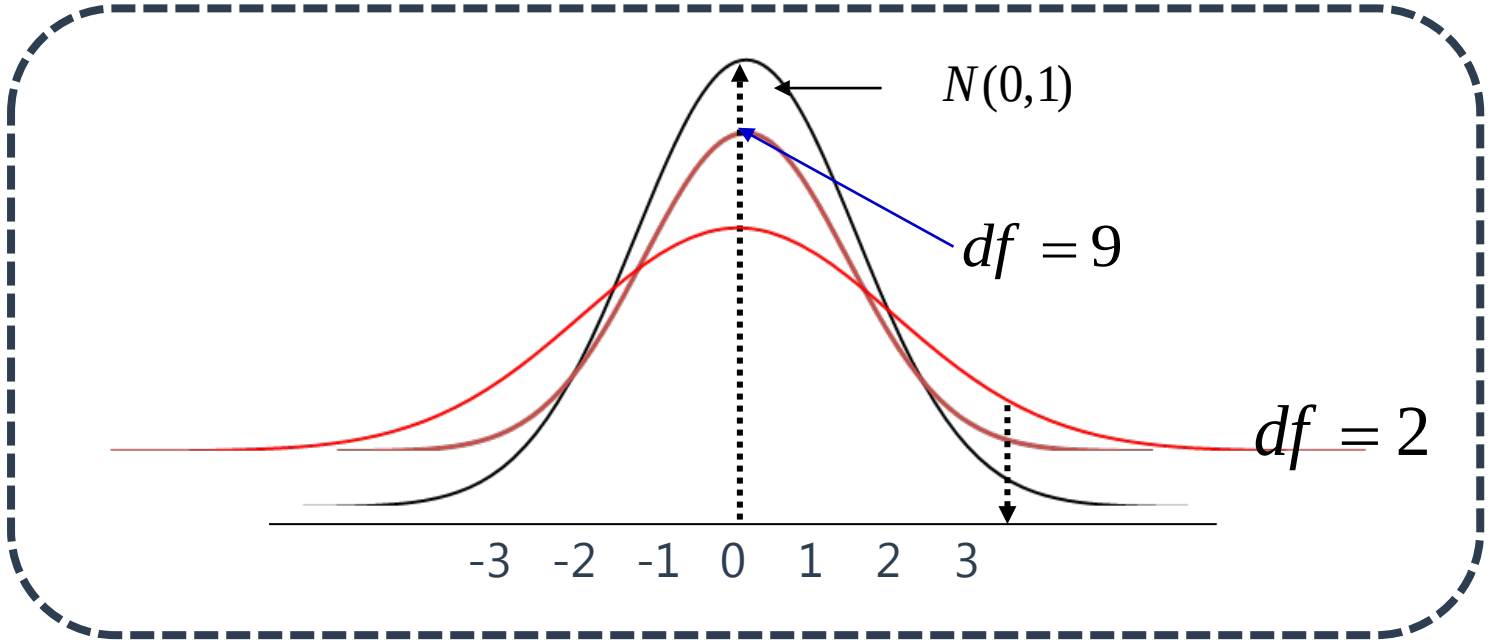
- **Example**

  - $Test\ Statistic = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$

  - $Test\ Statistic = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$

Student's t-distribution

- Degree of freedom=n-1
- Symmetrical
- Less peaked than a normal distribution ("fatter tails")
- Student's t-distribution converges to the standard normal distribution as degrees of freedom goes to infinity.



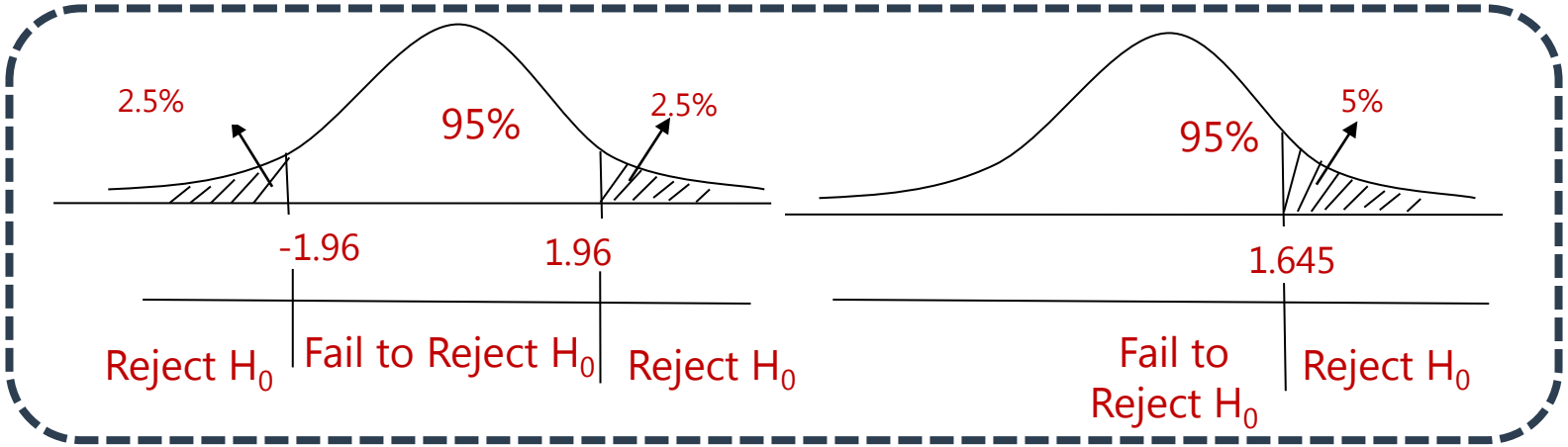
Hypothesis Testing

- **Step three: Specify the level of significance**
  - Critical value (关键值，实际就是分位数）
    - Found in the Z, T, Chi Square or F distribution tables not calculated by us
    - Under given one tailed or two tailed assumption, critical value is determined solely by the significance level.

- **Step four: state a decision rule**
  - Two tailed or one tailed test?
    Significance Level?
    Reject region? Critical Value under the condition
    Compare the Test Statistic and Critical Value

Hypothesis Testing

- **Step five: collect data and calculate the test statistic (critical value method)**
  - Find reject region with critical value;
  - **Reject** H_0 if | test statistic | > critical value; **fail to reject** H_0 if | test statistic | < critical value.



- **Step six: make a decision**
  - **cannot say** "accept the null hypothesis", only can say "cannot reject"
  - ***** is significantly different from ******
  - ***** is not significantly different from ******

Example

Two-tailed test

- You are analyzing Sendar Equity Fund, a midcap growth fund that has been in existence for 20 months with following data：
  - Mean monthly return = 1.50%
  - Sample standard deviation of monthly returns = 3.60%.
- Given its level of systematic risk and according to a pricing model, this mutual fund was expected to have earned a 1.10 percent mean monthly return during that time period. Assuming returns are normally distributed, are the actual results consistent with population mean monthly return of 1.10 percent at the 5% level of significance?

Example

Two-tailed test

- Correct Answer:
  - **Step 1:** We have a "not equal to" alternative hypothesis, where

$$H_0 : \mu = 1.10 \; versus \; H_a : \mu \neq 1.10$$

  - **Step 2 and 5:** Because of the normal population with unknown variance and a small sample, we use a **t-test** with 20-1=19 degree of freedom and calculate the test statistics t=(1.1%-1.5%)/(3.6%/$19^{0.5}$)=0.49
  - **Step 3:** Because this is a two-tailed test with 5% level of significance, with 19 df, the desired probability in each tail would be p =5%/2= 2.5%, giving the critical value equal to 2.093.
  - **Step 4:** Decision rule: reject the null if t>2.093 or t<-2.093.
  - **Step 6:** Because 0.497 does not satisfy either t>2.093 or t<-2.093, we cannot reject the null hypothesis.

Example

Two-tailed test

Table of the Student's t-Distribution (One-Tailed Probabilities)

df	p = 0.10	p = 0.05	**p = 0.025**	p = 0.01	p = 0.005
1	3.078	6.314	**12.706**	31.821	63.657
2	1.886	2.920	**4.303**	6.965	9.925
3	1.638	2.353	**3.182**	4.541	5.841
4	1.533	2.132	**2.776**	3.747	4.604
5	1.476	2.015	**2.571**	3.365	4.032
6	1.440	1.943	**2.447**	3.143	3.707
… …	… …	… …	**… …**	… …	… …
16	1.337	1.746	**2.120**	2.583	2.921
17	1.333	1.740	**2.110**	2.567	2.898
18	1.330	1.743	**2.101**	2.552	2.878
19	**1.328**	**1.729**	**2.093**	**2.539**	**2.861**
20	1.325	1.725	**2.086**	2.528	2.845

Example

One-tailed test

- Fashion Designs, a supplier of casual clothing to retail chains, is concerned about a possible slowdown in payments from its customers. The controller's office measures the rate of payment by the average number of days in receivables. Fashion Designs has generally maintained an average of 45 days in receivables. Because it would be too costly to analyze all of the company's receivables frequently, the controller's office uses sampling to track customers' payment rates. A random sample of 50 accounts shows a mean number of days in receivables of 49 with a standard deviation of 8 days. Determine whether the null hypothesis is rejected or not rejected at 10% level of significance.

Example

One-tailed test

- Correct Answer:
  - **Step 1:** The condition we want to reject is that the average number of days in receivables hasn't increased relative to the historical rate of 45 days, which suggests a "greater than" alternative hypothesis, where $H_0 : \mu \leq 45 \; versus \; H_a : \mu > 45$
  - **Step 2 and 5**: Because of the unknown population variance and a large sample, we can use either a z-test or a t-test with df=50-1=49. We choose t-test here and calculate the test statistics as

$$t = \frac{49 - 45}{8 / \sqrt{50}} = 3.536$$

  - **Step 3**: The critical value is found across the row for degrees of freedom of 49. To find the one-tailed rejection point for a 10% significance level, we use the 10% column and the value is 1.299.
  - **Step 4**: Decision rule: reject the null if t > 1.299.
  - **Step 6**: Because 3.536 >1.299, the null hypothesis is rejected at the 10% level, which gives us confidence that the mean has increased above 45 days.

Example

- An analyst conducts a two-tailed test to determine if earnings estimates are significantly different from reported earnings. The sample size was over 50. The computed Z-statistic is 1.58. At a 5 percent significance level, which of the following statements is TRUE?

  A. You cannot determine what to do with the information given.

  B. Fail to reject the null hypothesis and conclude that the earnings estimates are not significantly different from reported earnings.

  C. Both the null and the alternative are significant.

- **Correct Answer: B.**

Example

- An analyst is testing whether there are positive risk-adjusted returns to a trading strategy. He collects a sample and tests the hypotheses of H0: μ ≤ 0% versus Ha: μ > 0%, where μ is the population mean risk-adjusted return. The mean risk adjusted return for the sample is 0.7%. The calculated t-statistic is 2.428, and the critical t-value is 2.345. He estimates that the transaction costs are 0.3%. The results are most likely:

A. statistically and economically significant.

B. statistically significant but not economically significant.

C. economically significant but not statistically significant.

- **Correct Answer:**
  - A is correct. The results indicate that the mean risk-adjusted return is greater than 0% because the calculated test statistic of 2.428 is greater than the critical value of 2.245. The results are also economically significant because the risk adjusted return exceeds the transaction cost associated with this strategy by 0.4% (= 0.7 − 0.3).

P-value

- **P-value Method**
  - The **p-value** (P↓, easier to reject H_0)
    - the area in the probability distribution outside the calculated test statistic
    - the smallest level of significance at which the null hypothesis can be reject
  - p-value < α: reject H_0; p-value > α: do not reject H_0.

Example

- The p-value for a two-tailed test of sample mean is 1.68%. Which of the following is true?
  - A. We can reject the null with 95% confidence
  - B. We can reject the null with 99% confidence
  - C. the largest probability of rejecting the null hypothesis is 1.68%

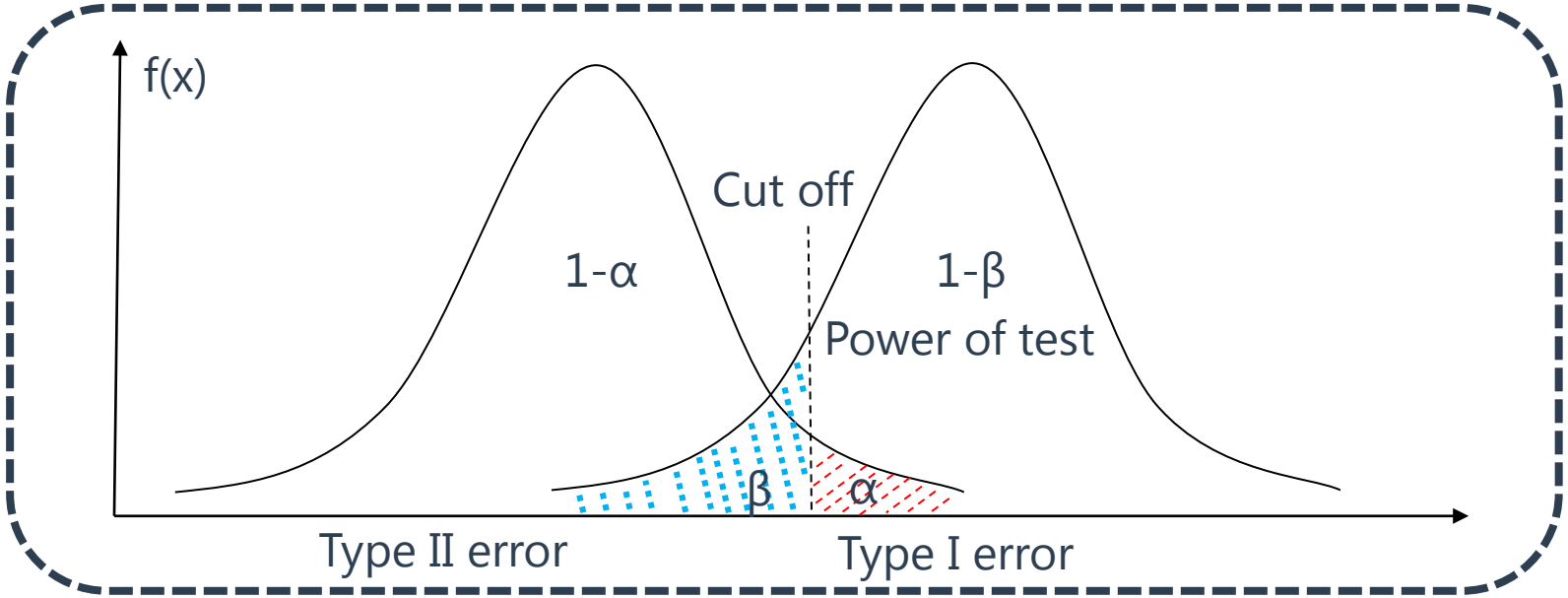
- Correct Answer: A

Type I Error or Type II Error

- **Type I error: 拒真，reject the null hypothesis when it's actually true**
  - Significance level (α): the probability of making a Type I error
  - Significance level =P(Type I error)=P(Reject | H_0√)
- **Type II error: 取伪，fail to reject the null hypothesis when it's actually false**
  - P(Type II error)=P(Fail to reject | H_0×)=Beta
  - **Power of a test:** the probability of **correctly** rejecting the null hypothesis (the probability of rejecting the null when it is false)
  - **Power of a test** = 1−P(Type II error)= P(Reject | H_0×)

Type I Error or Type II Error

- **How do we find the probability of type II error "Beta"?**



- Define a specific value for H_1 (Without this, we cannot calculate "Beta".
- Based the value of alpha, find the range of values outside the critical region of the test. (If your test statistic has been standardized, the range of values must be de-standardized.)
- Find the value of "cut off" and the "Beta", assuming H_1 is true. $P(H_0\surd \mid H_0X) = P(Ha\ X \mid Ha\surd)$
- power of test=1-Beta=$P(Ha\surd \mid Ha\surd)$=1- $P(H_0\surd \mid H_0X)$ =1-$P(Ha\times \mid Ha\surd)$

Type I Error or Type II Error

Decision	True condition	
	H_0 is false	**H_0 is true**
Reject H_0	Correct Decision Power of test = 1- P (Type Ⅱ error)	Incorrect Decision Significance level =P (Type I error)
Do not reject H_0	Incorrect Decision Type Ⅱ error	Correct Decision

- With other conditions unchanged, either error probability arises at the cost of the other error probability decreasing.
- How to reduce both errors? Increase the Sample Size.

Example

- If the sample size increases, the probability of get the Type I and Type II error will

	Type I	Type II
A.	increase	increase
B.	not change	not change
C.	decrease	decrease

- **Correct Answer: C**

Example

- All else equal, is specifying a larger significance level in a hypothesis test likely to increase the probability of a:

	type I error?	type II error?
A.	No	No
B.	No	Yes
C.	Yes	No

- **Correct Answer: C.**

- What is the definition of the power of test? Power of test is the probability to:

  A. Reject the true null hypothesis while it is true
  B. Reject the false null hypothesis while it is indeed false
  C. Can not reject the true hypothesis

- **Correct Answer: B.**

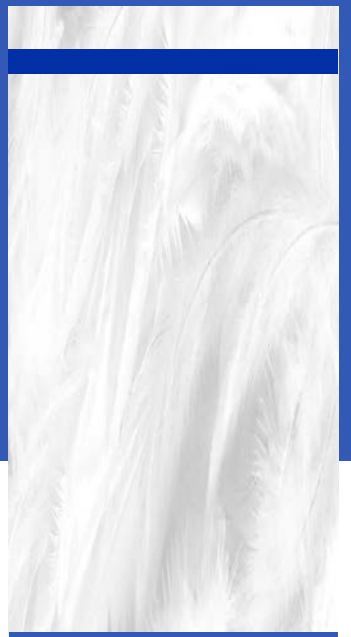
Summary

Hypothesis Testing

Hypothesis tests for finance

Tests of return and risk in finance

- Construct hypothesis tests and determine their statistical significance, the associated Type I and Type II errors, and power of the test given a significance level



Summary of Hypothesis Testing

Test type	Assumptions	H_0	Test-statistic	Critical value and degree of freedom
Mean hypothesis testing	Normally distributed population, known population variance	$\mu=0$	$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	N(0,1)
	Normally distributed population, unknown population variance	$\mu=0$	$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	t(n-1)
	Independent populations, unknown population variances **assumed equal**	$\mu_1-\mu_2=0$	$t=\frac{(\bar{x_1}-\bar{x_2})-(\mu_1-\mu_2)}{\sqrt{s_p^2/n_1+s_p^2/n_2}}$ *where* s_p^2 *is a pooled estimator* $s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$t(n_1+n_2-2)$
	Samples not independent (paired comparisons test)	$\mu_d=0$	$t=\bar{d}/s_{\bar{d}}$ $s_{\bar{d}}=s_d/\sqrt{n}$	t(n-1)

Example

- An analyst collects the following data related to paired observations for Sample A and Sample B. Assume that both samples are drawn from normally distributed populations and that the population variances are not known.

Paired Observation	Sample A Value	Sample B Value
1	25	18
2	12	9
3	-5	-8
4	6	3
5	-8	1

The *t*-statistic to test the hypothesis that the mean difference is equal to zero is *closest* to:

A. 0.23
B. 0.27
C. 0.52

➢ **Correct Answer: C**

Example

- Smith wants to know whether the mean returns of two stocks are the same. If the two normally distributed stock returns are dependent, the appropriate type of test and test statistic are:

A. Difference in means test, t-statistic

B. Paired comparisons test, t-statistic

C. Difference in means test, F-statistic

- **Correct Answer: B.**

Summary of Hypothesis Testing

Test type	Assumptions	H_0	Test-statistic	Critical value and degree of freedom
Variance hypothesis testing	Normally distributed population	$\sigma^2=\sigma_0^2$	$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2(n-1)$
	Two independent normally distributed populations	$\sigma_1^2=\sigma_2^2$	$F=s_1^2/s_2^2$	$F(n_1-1,n_2-1)$

Example
Chi-square test for a single population variance

- ABC Equity Fund has been in existence for 24 months. During this period, it has achieved a mean monthly return of 1.50 percent with a sample standard deviation of monthly returns of 3.70 percent. An analyst now wants to test a claim that the particular investment disciplines followed by ABC result in a standard deviation of monthly returns of less than 4 percent at the 0.01 level of significance.

- Correct Answer:
  - **Step 1**: We have a "less than" alternative hypothesis, where σ is the underlying standard deviation of return on ABC Equity Fund. Being careful to square standard deviation to obtain a test in terms of variance, the hypotheses are

    H_0: $\sigma^2 \geq 16.0$ versus H_a: $\sigma^2 < 16.0$.

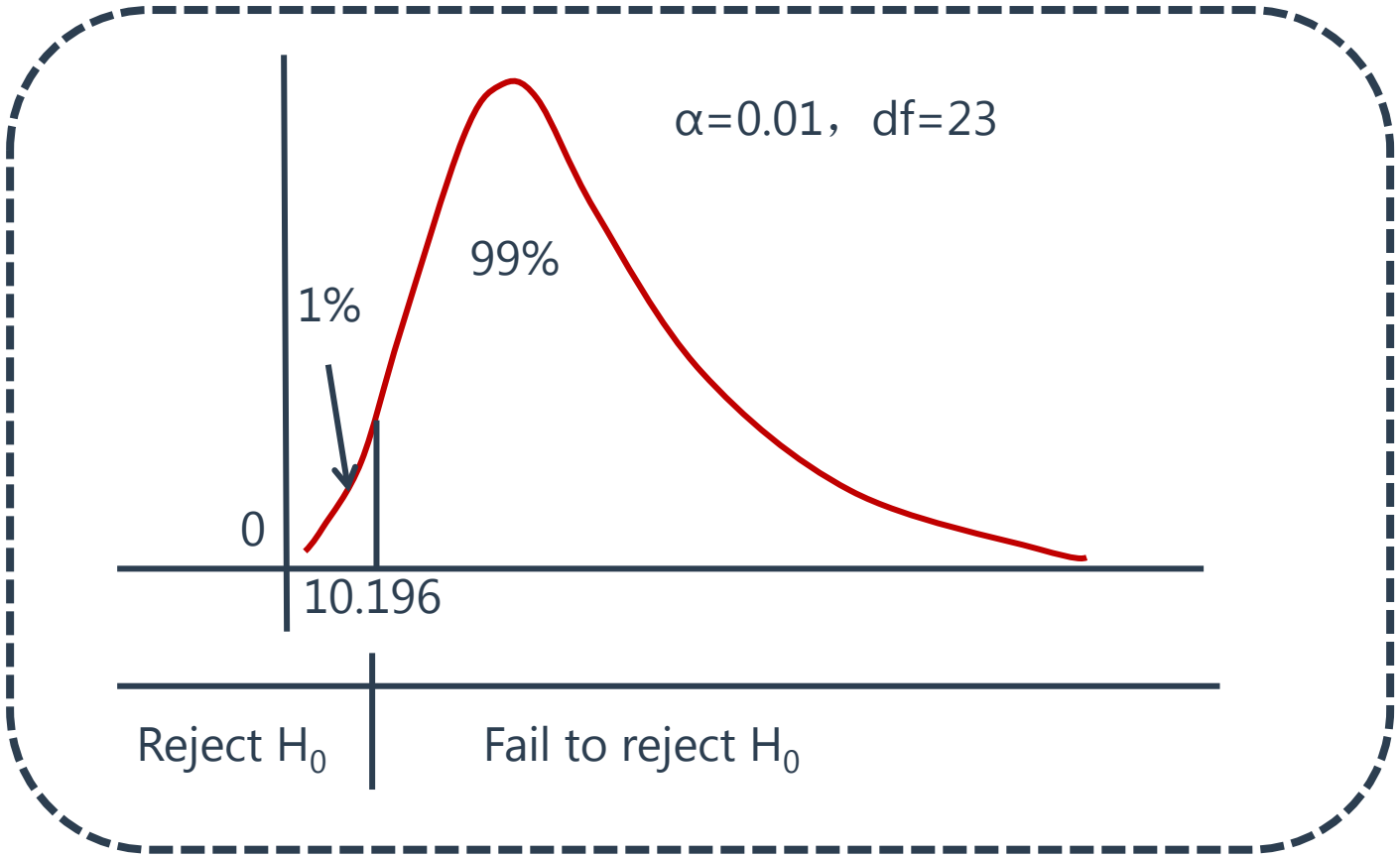
  - **Step 2 and 5**: The test statistic is χ^2 with 24 − 1 = 23 degrees of freedom.

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{23 \times 3.70^2}{4^2} = 19.68$$

Example

Chi-square test for a single population variance

- Decision Rule for a One-Tailed Chi-Square Test of a Single Population Variance



Example

Chi-square test for a single population variance

- **Step 3**: The lower 0.01 rejection point ($\alpha=0.01$) is found on the line for df = 23, under the 0.99 column (99 percent probability in the right tail, to give 0.99 probability of getting a test statistic this large or larger). The rejection point is 10.196.
- **Step 4**: We will reject the null if we find that χ^2 is less than 10.196.
- **Step 6**: Because 19.68 (the calculated value of the test statistic) is not less than 10.196, we do not reject the null hypothesis. We cannot conclude that ABC's investment disciplines result in a standard deviation of monthly returns of less than 4 percent.

Example

F-test for equal variances

- An analyst are investigating whether the population variance of returns on the KOSPI Index of the South Korean stock market changed subsequent to the global financial crisis that peaked in 2008. For this investigation, you are considering 2004 to 2006 as the pre-crisis period and 2010 to 2012 as the post-crisis period. You gather the data in Table for 156 weeks of returns during 2004 to 2006 and 156 weeks of returns during 2010 to 2012. You have specified a 0.01 level of significance.

KOSPI Index Returns and Variance before and after the Global Financial Crisis of the Late 2000s

	n	Mean Weekly Return (%)	Variance of Returns
Before crisis: 2004 to 2006	156	0.358	7.240
After crisis: 2010 to 2012	156	0.110	6.269

Example

F-test for equal variances

- **Step 1**: We have a "not equal to" alternative hypothesis:

$$H_0 : \sigma^2_{Before} = \sigma^2_{After} \quad \text{versus} \quad H_a : \sigma^2_{Before} \neq \sigma^2_{After}$$

- **Step 2 and 5**: Select the appropriate test statistic. For tests of difference between variances, the appropriate test statistic is $F = s_1^2 / s_2^2$ with 156 − 1 = 155 numerator and denominator degrees of freedom.

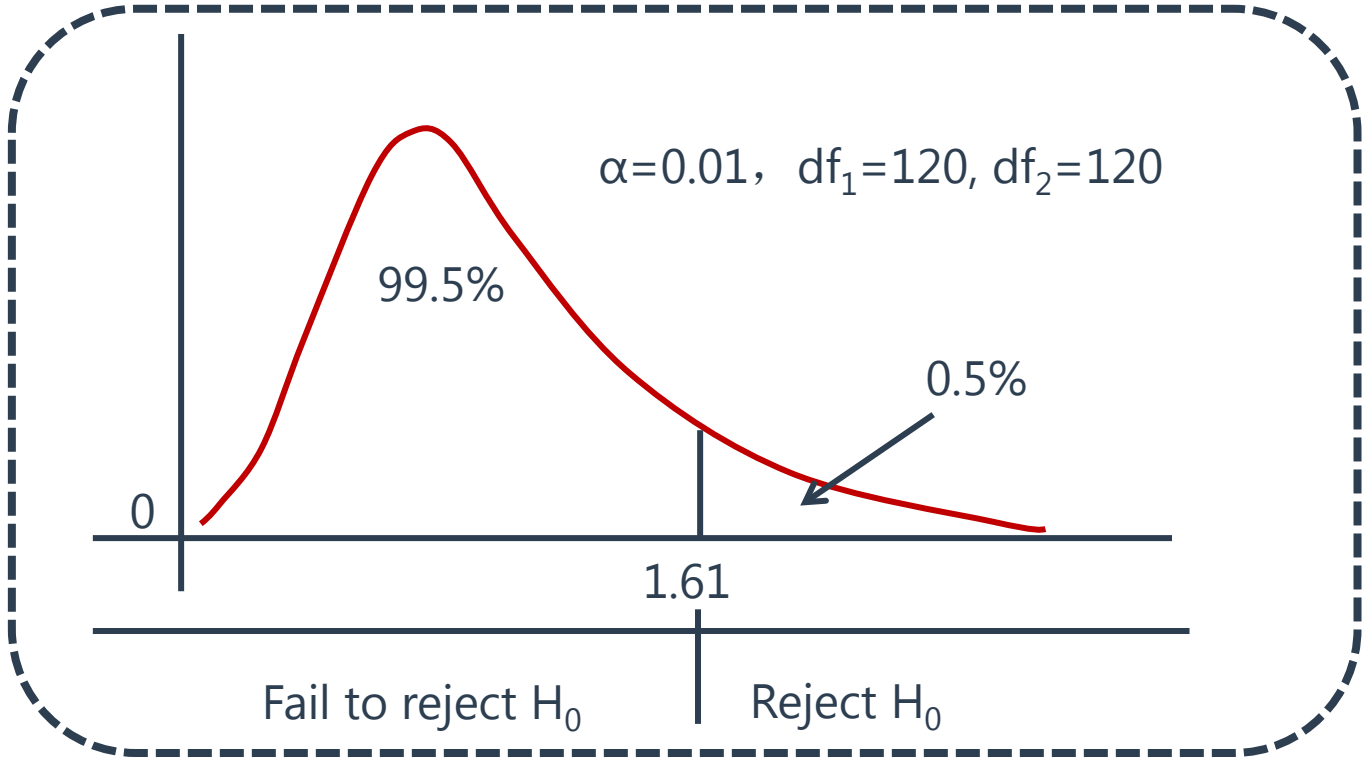
$$F = s_1^2 / s_2^2 = 7.240 / 6.269 = 1.155$$

- **Step 3**: This is a two-tailed test, we use F-tables for the 0.005 level (= 0.01/2) to give a 0.01 significance level （α=0.01）. In the tables in the back of the volume, the closest value to 155 degrees of freedom is 120 degrees of freedom. At the 0.01 level, the rejection point is 1.61.

Example

F-test for equal variances

- **Decision rule for F-test**



- **Step 4**: We will reject the null if we find that χ^2 is more than 1.61.
- **Step 6**: Because 1.155 is less than the critical value 1.61, we cannot reject the null hypothesis that the population variance of returns is the same in the pre- and post-global financial crisis periods.

Example

- Which of the following is *most appropriate* to test the equality of the variances of two normally distributed populations?

  A. F-test

  B. T-test

  C. χ^2-test

- **Correct Answer: A.**

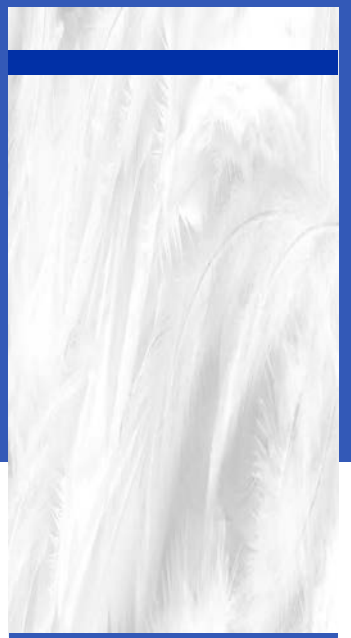
Summary

Hypothesis Testing

Tests of return and risk in finance

Parametric versus nonparametric tests

- Compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test



Hypothesis Testing

- **Parametric tests**
  - based on specific distributional assumptions for the population
  - concerning a parameter of population.
  - For example, t-test.
- **Nonparametric tests**
  - a nonparametric test either is not concerned with a parameter or makes minimal assumptions about the population from which the sample comes.
  - **Nonparametric tests are used:**
    - **when data do not meet distributional assumptions.**
      - Example: hypothesis test of the mean value for a variable, but the distribution of the variable is not normal and the sample size is small so that neither the t-test nor the z-test are appropriate.
    - **when there are outliers.**
    - **when data are given in ranks.**
    - **when the hypothesis we are addressing does not concern a parameter.**

Example

Spearman rank correlation coefficient

- You are interested in whether excess risk-adjusted return (alpha) is correlated with mutual fund expense ratios for US large-cap growth funds. The following table presents the sample.

Mutual Fund	Alpha	Expense Ratio
1	−0.52	1.34
2	−0.13	0.40
3	−0.50	1.90
4	−1.01	1.50
5	−0.26	1.35
6	−0.89	0.50
7	−0.42	1.00
8	−0.23	1.50
9	−0.60	1.45

- **A** Formulate null and alternative hypotheses consistent with the verbal description of the research goal.
- **B** Identify and justify the test statistic for conducting a test of the hypotheses in Part A.
- **C** Determine whether to reject the null hypothesis at the 0.05 level of significance if the critical values are ±2.306.

Example

Spearman rank correlation coefficient

- **Correct Answer:**
  - **A** We have a “not equal to” alternative hypothesis:
    - *H*0: ρ = 0 versus *Ha*: ρ ≠ 0
  - **B** Mutual fund expense ratios are bounded from above and below; in practice, there is at least a lower bound on alpha (as any return cannot be less than −100%), and expense ratios cannot be negative. These variables may not be normally distributed, and the assumptions of a parametric test are not likely to be fulfilled. Thus, a nonparametric test appears to be appropriate. We would use the nonparametric Spearman rank correlation coefficient to conduct the test:

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

with the *t*-distributed test statistic of $t = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}}$

Example

Spearman rank correlation coefficient

- **Correct Answer:**
  - **C** The calculation of the Spearman rank correlation coefficient is given in the following table.

Mutual Fund	Alpha	Expense Ratio	Rank by Alpha	Rank by Expense Ratio	Difference in Rank	Difference Squared
1	−0.52	1.34	6	6	0	0
2	−0.13	0.40	1	9	−8	64
3	−0.50	1.90	5	1	4	16
4	−1.01	1.50	9	2	7	49
5	−0.26	1.35	3	5	−2	4
6	−0.89	0.50	8	8	0	0
7	−0.42	1.00	4	7	−3	9
8	−0.23	1.50	2	2	0	0
9	−0.60	1.45	7	4	3	9
						151

Example

Spearman rank correlation coefficient

- **Correct Answer:**
  - **C**

$$r_S = 1 - \frac{6(151)}{9(80)} = -0.25833$$

The calculated test statistic, using the *t*-distributed test statistic

$$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{-0.25833\sqrt{7}}{\sqrt{1 - 0.066736}} = \frac{-0.683486}{0.9332638} = -0.7075$$

On the basis of this value falling within the range of +2.306, we fail to reject the null hypothesis that the Spearman rank correlation coefficient is zero.

Summary

Hypothesis Testing

Parametric versus Nonparametric Tests

Summary

Module: Hypothesis Testing

Hypothesis tests for finance

Tests of return and risk in finance

Parametric versus Nonparametric Tests

Module

Parametric and Non-Parametric Tests of Independence

1. Tests concerning correlation
2. Tests of Independence Using Contingency Table Data

Tests concerning correlation

- Explain parametric and nonparametric tests of the hypothesis that the population correlation coefficient equals zero, and determine whether the hypothesis is rejected at a given level of significance

Significance test of the correlation

Test type	Assumptions	H_0	Test-statistic	Critical value and degree of freedom
Correlation	Both of the variables are normally distributed.	ρ = 0	$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}}$	t(n-2)

Parametric Test of a Correlation

- The **parametric pairwise correlation coefficient** is often referred to as **Pearson correlation**, the **bivariate correlation**, or simply **the correlation**.
- The process of hypothesis testing about whether the correlation between the population of two random variables is significantly different from zero.
  - Step 1: H_0: ρ=0; H_a: ρ≠0 (Two-tailed test)
  - Step 2 and 5: Calculate the test statistic:

$$t = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}}, \text{ df=n-2}$$

  - Step 3: Specify the level of significance.
  - Step 4: Decision rule: reject H_0 if $|t| > +t_{critical}$
  - Step 6: Draw a conclusion that the correlation between the population of two variables is **significantly different from zero** if **H_0 is rejected.**

Example

- The covariance between X and Y is 16. The standard deviation of X is 4 and the standard deviation of Y is 8. The sample size is 20. Test the significance of the correlation coefficient at the 5% significance level.

- **Correct Answer:**
  - The sample correlation coefficient r = 16/(4×8) = 0.5. The t-statistic can be computed as:

$$t = 0.5 \times \frac{\sqrt{20-2}}{\sqrt{1-0.25}} = 2.45$$

The critical t-value for α=5%, two-tailed test with df=18 is 2.101.

Since the test statistic of 2.45 is larger than the critical value of 2.101, we have sufficient evidence to reject the null hypothesis. So we can say that the correlation coefficient between X and Y is significantly different from zero.

Example

Table of the Student's t-Distribution (One-Tailed Probabilities)

df	p = 0.10	p = 0.05	**p = 0.025**	p = 0.01	p = 0.005
1	3.078	6.314	**12.706**	31.821	63.657
2	1.886	2.920	**4.303**	6.965	9.925
3	1.638	2.353	**3.182**	4.541	5.841
4	1.533	2.132	**2.776**	3.747	4.604
5	1.476	2.015	**2.571**	3.365	4.032
6	1.440	1.943	**2.447**	3.143	3.707
… …	… …	… …	**… …**	… …	… …
16	1.337	1.746	**2.120**	2.583	2.921
17	1.333	1.740	**2.110**	2.567	2.898
18	**1.330**	**1.743**	**2.101**	**2.552**	**2.878**
19	1.328	1.729	**2.093**	2.539	2.861
20	1.325	1.725	**2.086**	2.528	2.845

Example

- Jill Batten is analyzing how the returns on the stock of Stellar Energy Corp. are related with the previous month's percent change in the US Consumer Price Index for Energy (CPIENG). Based on 248 observations, she has computed the sample correlation between the Stellar and CPIENG variables to be −0.1452. She also wants to determine whether the sample correlation is statistically significant. The critical value for the test statistic at the 0.05 level of significance is approximately 1.96. Batten should conclude that the statistical relationship between Stellar and CPIENG is:

A. significant, because the calculated test statistic has a lower absolute value than the critical value for the test statistic.

B. significant, because the calculated test statistic has a higher absolute value than the critical value for the test statistic.

C. not significant, because the calculated test statistic has a higher absolute value than the critical value for the test statistic.

Example

- **Correct Answer: B.**
  - The sample correlation coefficient r = -0.1452. The t-statistic can be computed as:

$$t = (-0.1452) \times \frac{\sqrt{248-2}}{\sqrt{1-(-0.1452)^2}} = -2.3018$$

  - The critical value for the test statistic at the 0.05 level of significance is approximately 1.96. Since the test statistic of -2.3018 is less than the critical value of -1.96, we have sufficient evidence to reject the null hypothesis. So we can say that the correlation coefficient is significantly different from zero.

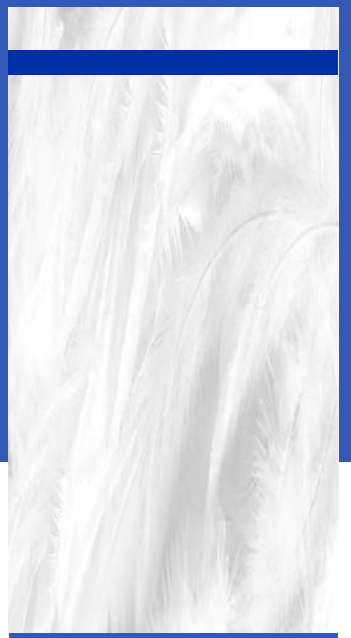
Summary

Hypothesis Testing

Tests concerning correlation

Tests of independence using contingency table data

- Explain tests of independence based on contingency table data



Tests of independence

- Test whether there is **a relationship between the size and investment type**, we can perform a test of independence using a nonparametric test statistic that is chi- square distributed

$$\chi^2 = \sum_{i=1}^{m} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

  - m = the number of cells in the table, which is the number of groups in the first class multiplied by the number of groups in the second class.
  - O_{ij} = the number of observations in each cell of row i and column j.
  - E_{ij} = the expected number of observations in each cell of row i and column j, assuming independence.
  - degrees of freedom is (r − 1)(c − 1), where r is the number of rows and c is the number of columns.

Observed vs. Expected value

- Observed Values

	Low Risk	High Risk	
Growth	73	26	99
Value	183	33	216
	256	59	315

	Low Risk	High Risk	
Growth	74%	26%	100%
Value	85%	15%	100%

- Expected Value$_{i,j}$ E_{ij} = (Total Row i × Total Column j)/Overall Total
  - E.g. Expected value for Growth/Low Risk is: (99 × 256)/ 315 = 80.46

	Low Risk	High Risk	
Growth	80.457	18.543	99
Value	175.543	40.457	216
	256	59	315

	Low Risk	High Risk	
Growth	81%	19%	100%
Value	81%	19%	100%

Example

Tests of independence

- This table is referred to as a **contingency table** (or a **two-way table**, because there are two classifications, or classes—size and investment type). If we want to test whether there is a relationship between the size and investment type of 1,594 exchange traded funds (ETFs), we can perform a test of independence using a nonparametric test statistic that is chi- square distributed.

Investment Type	Size Based on Market Capitalization			
	Small	Medium	Large	Total
Value	50	110	343	503
Growth	42	122	202	366
Blend	56	149	520	725
Total	148	381	1065	1594

Example

Tests of independence

- **Correct Answer:**
  - **Step 1:** H_0: ETF size and investment type are not related, so these classifications are independent versus H_a : ETF size and investment type are related, so these classifications are not independent.
  - **Step 2 and 5 :** With (3 − 1) × (3 − 1) = 4 degrees of freedom. $E_{11} = \frac{(\text{Total row } i)\times(\text{Total column } j)}{\text{Overall total}} = \frac{503\times148}{1594} = 46.703$, repeat this calculation for each combination of size and investment type (m=3×3=9 pairs) to arrive at the expected frequencies. Next calculate the test statistic χ^2=32.08025.
  - **Step 3:** Because this is a one-sided test with a 5% level of significance, the critical value is 9.4877.
  - **Step 4:** State a decision rule: reject the null hypothesis if χ^2>9.4877.
  - **Step 6:** Because 32.08025 >9.4877, there is sufficient evidence to conclude that ETF size and investment type are related(not independent).

Summary

Hypothesis Testing

Tests of Independence Using Contingency Table Data

Summary

Module: Hypothesis Testing

Tests concerning correlation

Tests of Independence Using Contingency Table Data

Module

Simple Linear Regression

1. Basics of simple linear regression
2. Estimate
3. Hypothesis testing
4. Estimate of Y
5. Forms of Simple Linear Regression

Basics of Simple Linear Regression

- Describe a simple linear regression model
- Explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated

Simple Linear Regression

- The **simple linear regression** (is often referred to as **ordinary least squares** (OLS) regression)

$$Y_i = b_0 + b_1 X_i + \varepsilon_i \, , i = 1, \dots, n$$

  - The goal is to fit a line to the observations on *Y* and *X* to minimize the squared deviations from the line; this is the least squares criterion—hence, the name **least squares regression**.

- **Interpretation of the parameters**

  - **The dependent variable, Y** is the variable whose variation about its mean is to be ***explained*** by the regression.
  - **The independent variable, X** is the variable used to ***explain*** the dependent variable in a regression.
  - **Regression coefficients, b_0** is intercept term of the regression, **b_1** is slope coefficient of the regression, regression coefficient.
  - **The error term (residual term), ε_i** is the portion of the dependent variable that is not explained by the independent variable(s) in the regression.

Basic concepts

- **Linear regression** with one regressor
  - use linear regression to summarize the **linear** relationship;
  - use one variable to make predictions about another.
- **Cross-Sectional** versus **Time-Series Regressions**
  - A cross-sectional regression involves many observations of X and Y (**cross-sectional data**) for the same time period.
  - **Time-series data** use many observations from different time periods for the same company, asset class, investment fund, country, or other entity, depending on the regression model.

Dummy variable

- **Dummy variable (indicator variable)**
  - Qualitative independent variable (X)
  - Distinguish among n categories, need n-1 dummy variables.
- **Example: Does gender have an impact on a mutual fund's performance?**
  - There are two types of gender (male and female). So, only one (=2-1) dummy variable "X" is required.
  - $X=\begin{cases}1 & female\\ 0 & male\end{cases}$
  - Y=Return of portfolio (R_p), that indicates the performance of the mutual fund manager
- $Y_i = b_0 + b_1 X_i + \varepsilon_i\,, i = 1, \dots, n$
  - $\hat{b}_0$: X=0, $\hat{Y}_0$ indicate the performance of a manager (male), $\hat{Y}_0 = \hat{b}_0$
  - $\hat{b}_1 = \hat{Y}_1$（X=1, $\hat{Y}_1$ indicate the performance of a manager (female)）- $\hat{b}_0$

Assumptions of the Linear Regression

- **Linearity:** The relationship between the dependent variable, *Y*, and the independent variable, *X* is **linear** in the parameters b_0 **and** b_1**.**
  - $Y_i = b_0 e^{b_1 X_i} + \varepsilon_i$ is nonlinear in b_1, so we could not apply the linear regression model to it.
  - Even if the dependent variable is nonlinear, $Y_i = b_0 + b_1 x_i^2 + \varepsilon_i$, however, linear regression can still be used to estimate.
- The independent variable, X, is **not random**, with the exception that X is random but also **uncorrelated with the error term**.
- The expected value of the error term is zero (i.e., $E(\varepsilon_i)=0$)
- **Homoskedasticity:** The variance of the error term is **constant**. If not, this refers to heteroskedasticity
- **Independence:** The **error term is uncorrelated** across observations
- **Normality:** The error term is normally distributed.

Example

- An analyst has estimated a model that regresses a company's return on equity (ROE) against its growth opportunities (GO), defined as the company's three year compounded annual growth rate in sales, over 20 years and produces the following estimated simple linear regression:

$$ROE_i = 4 + 1.8GO_i + \varepsilon_i$$

Both variables are stated in percentages, so a GO observation of 5% is included as 5.

- **1** The predicted value of the company's ROE if its GO is 10% is closest to:

A. 1.8%.

B. 15.8%.

C. 22.0%.

- **2** The change in ROE for a change in GO from 5% to 6% is closest to:

A. 1.8%.

B. 4.0%.

C. 5.8%.

Example

- **3** The residual in the case of a GO of 8% and an observed ROE of 21% is closest to:

A. −1.8%.

B. 2.6%.

C. 12.0%.

- **Correct Answer:**
  - **1** C is correct. The predicted value of ROE = 4 + (1.8 × 10) = 22.
  - **2** A is correct. The slope coefficient of 1.8 is the expected change in the dependent variable (ROE) for a one-unit change in the independent variable (GO).
  - **3** B is correct. The predicted value is ROE = 4 + (1.8 × 8) = 18.4. The observed value of ROE is 21, so the residual is 2.6 = 21.0 − 18.4.

Summary

Simple Linear Regression

Basics of Simple Linear Regression

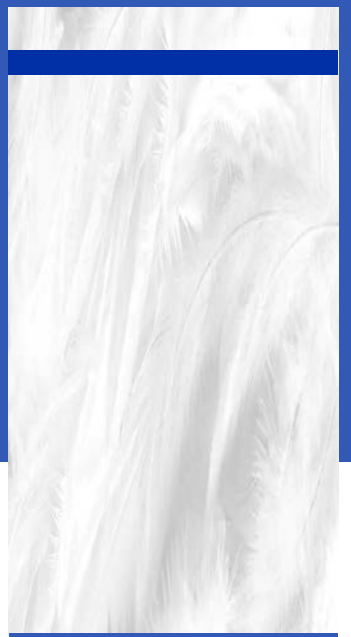
Interpret parameters of simple linear regression

Explain dummy variables

Explain assumptions of simple linear regression

Estimate of Regression Coefficients

- Describe how the least squares criterion is used to estimate regression coefficients, and the interpretation of these coefficients



Point Estimate

- **Point estimate:** $\hat{b}_1 = b_1 \quad \hat{b}_0 = b_0$
- **Calculation of $\hat{b}_1$ and $\hat{b}_0$**
  - **Ordinary least squares (OLS):** Minimize the sum of squared vertical distances between the observations and the regression line (also called residuals or error terms).
  - $\hat{b}_1 = \frac{Cov(X,Y)}{Var(X)}; \hat{b}_0 = \bar{Y} - \hat{b}_1\bar{X}$.
- **Interpretation**
  - The **estimated slope coefficient** ($\hat{b}_1$) defines the sensitivity of Y to a change in X.
  - The **estimated intercept coefficient** ($\hat{b}_0$) refers to the value of Y when X is equal to zero.
- **Contrast:** $\hat{b}_1$ and r
  - The sample correlation, r, is the ratio of the covariance to the product of the standard deviations:

$$r = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Example

- Bouvier Co. is a Canadian company that sells forestry products to several Pacific Rim customers. Bouvier's sales are very sensitive to exchange rates. The following table shows recent annual sales (in millions of Canadian dollars) and the average exchange rate for the year (expressed as the units of foreign currency needed to buy one Canadian dollar).

Year i	X_i = Exchange Rate	Y_i = Sales
1	0.40	20
2	0.36	25
3	0.42	16
4	0.31	30
5	0.33	35
6	0.34	30

- Calculate the intercept and coefficient for an estimated linear regression with the exchange rate as the independent variable and sales as the dependent variable.

Example

- The sample mean of the exchange rate is:

$$\bar{X} = \sum_{i=1}^{n} X_i / n = 2.16 / 6 = 0.36$$

- The sample mean of sales is:

$$\bar{Y} = \sum_{i=1}^{n} Y_i / n = 156 / 6 = 26$$

- We want to estimate a regression equation of the form $Y_i = b_0 + b_1 X_i + \varepsilon_i$. The estimates of the slope coefficient and the intercept are

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{-1.39}{0.009} = -154.44$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} = 26 - (-154.444)(0.36) = 26 + 55.6 = 81.6$$

- So the regression equation is $Y_i = 81.6 - 154.444 X_i$

Confidence Interval Estimate

- Regression coefficient confidence interval

$$\hat{b}_1 \pm t_c S_{\hat{b}_1}$$

  - $\hat{b}_1 = \frac{Cov(X,Y)}{Var(X)}$
  - $S_{\hat{b}_1}$ is the **standard error** of the estimated coefficient $\hat{b}_1$.
  - t_c（查表）df=n-2（e.g. n=20, alpha=5%)
  - Regression models with good fitness will lead to **smaller** standard error of an estimated coefficient $S_{\hat{b}_1}$ and **tighter** confidence intervals.

- **Application**: If the confidence interval with a given degree of confidence does **not** include the hypothesized value, the null is rejected, and the coefficient is said to be statistically different from hypothesized value.

Critical value: t-table

Appendix B Table of the Student's t-Distribution (One-Tailed Probabilities)

df	p = 0.10	p = 0.05	**p = 0.025**	p = 0.01	p = 0.005
1	3.078	6.314	**12.706**	31.821	63.657
2	1.886	2.920	**4.303**	6.965	9.925
3	1.638	2.353	**3.182**	4.541	5.841
4	1.533	2.132	**2.776**	3.747	4.604
5	1.476	2.015	**2.571**	3.365	4.032
6	1.440	1.943	**2.447**	3.143	3.707
… …	… …	… …	**… …**	… …	… …
16	1.337	1.746	**2.120**	2.583	2.921
17	1.333	1.740	**2.110**	2.567	2.898
18	**1.330**	**1.743**	**2.101**	**2.552**	**2.878**
19	1.328	1.729	**2.093**	2.539	2.861
20	1.325	1.725	**2.086**	2.528	2.845

Summary

Simple Linear Regression

Estimate of regression coefficients

Calculate estimate of slope and intercept with a calculator

Calculate confidence interval of slope or intercept

Hypothesis Test

- Calculate and interpret measures of fit and formulate and evaluate tests of fit and of regression coefficients in a simple linear regression
- 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

Significance test for regression coefficient

- **Significance test for regression coefficient**
  - H_0: $b_1 = 0$
  - Test Statistic= $t = \frac{\hat{b}_1 - 0}{S_{\hat{b}_1}} = \frac{\hat{b}_1}{S_{\hat{b}_1}}$
  - t critical（查表）df=n-2
  - Decision rule: reject H_0, if | T.S. | > + t critical
  - Rejection of the null means that the slope coefficient is significantly different from zero.

Abbreviation table

Abbreviation	English	Chinese
$\hat{b}_1$	Sample/estimated slope	样本/（回归）估计的斜率
b_1	Population slope	总体斜率
$\hat{b}_0$	Sample/estimated intercept	样本/（回归）估计的截距
b_0	Population intercept	总体截距
ε_i	Error term/residual term	误差项/残差项
Y_i	Actual Y	真实Y
$\hat{Y}_1$	Estimated Y/Point estimate of Y	（直线回归）估计的Y/Y的点估计的值
TSS	Total sum of square/total variation	总平方和/总偏离
RSS	Sum of square of regression	回归的平方和（可以被回归解释的偏离）
SSE	Sum of square of error	误差（项）的平方和（不可以被回归解释的偏离）
$S_{\hat{b}_1}$	Standard error of $\hat{b}_1$	$\hat{b}_1$的标准误
T.S.	Test statistics	检验统计量
CV	Critical Value	关键值
df	Degree of freedom	自由度
n	Sample Size	样本容量

Hypothesis testing

- **Hypothesis testing about regression coefficient**
  - H_0: b_1= hypothesized value of b_1
  - Test Statistic:

$$t = \frac{\hat{b}_1 - hypothesized\ value\ of\ b_1}{S_{\hat{b}_1}}\text{ , df=n-2}$$

  - **Decision rule**: reject H_0 if | t | > + $t_{critical}$
  - Rejection of the null means that the slope coefficient is significantly different from the hypothesized value of b_1.

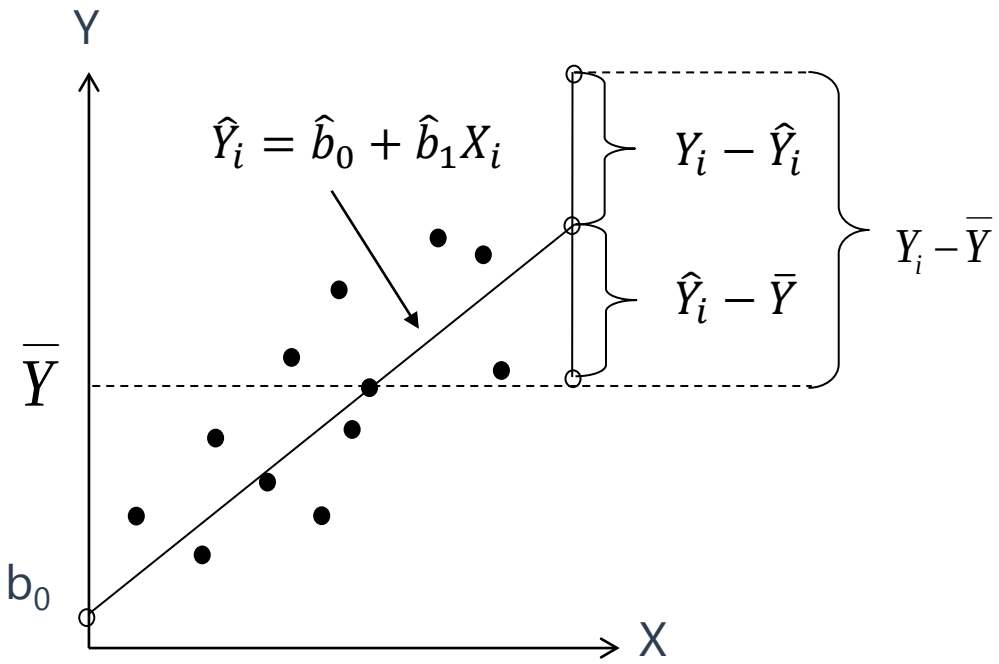
P-value

- **P-value Method**
  - H_0: $b_1 = 0$
  - The **p-value** is the smallest level of significance at which the null hypothesis can be reject.
  - p-value < α: reject H_0.
  - reject H_0 means the coefficient is significantly different from zero.

	Coefficient	t-statistic	p-value
Intercept	-0.5	-0.91	0.18
Slope	2	20.00	<0.001

Measure Fitness-ANOVA Table

- **Elements of ANOVA Table**
  - The sum of squared errors or residuals (SSE)= $\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$
  - The regression sum of squares (RSS)= $\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$
  - Total sum of squares (TSS)= $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$



Abbreviation table

Abbreviation	English	Chinese
$\hat{b}_1$	Sample/estimated slope	样本/（回归）估计的斜率
b_1	Population slope	总体斜率
$\hat{b}_0$	Sample/estimated intercept	样本/（回归）估计的截距
b_0	Population intercept	总体截距
ε_i	Error term/residual term	误差项/残差项
Y_i	Actual Y	真实Y
$\hat{Y}_1$	Estimated Y/Point estimate of Y	（直线回归）估计的Y/Y的点估计的值
$S_{\hat{b}_1}$	Standard error of $\hat{b}_1$	$\hat{b}_1$的标准误
T.S.	Test statistics	检验统计量
CV	Critical Value	关键值
df	Degree of freedom	自由度
n	Sample Size	样本容量
k	Number of X	（自变量）X的个数

Abbreviation table

Abbreviation	English	Chinese
ANOVA	Analysis of variance (Table)	方差分析表
SS	Sum of square/variation	平方和/偏离
TSS (SST)	Total sum of square/total variation	总平方和/总偏离
RSS (SSR)	Sum of square of regression	回归的平方和（可以被回归解释的偏离）
SSE (ESS)	Sum of square of error	误差（项）的平方和（不可以被回归解释的偏离）
MSS	Mean square	均方（=平方和/自由度，即方差）
MSR	Mean square regression	回归均方（方差）
MSE	Mean square error	误差均方（方差）
S_e (SEE)	Standard error of the estimate	误差项的标准差/（回归）估计的标准误
R^2	Coefficient of determination/R-square	判定系数/R平方
$r_{x,y}$	Correlation coefficient of X and Y	X和Y的相关系数

Measure Fitness-ANOVA Table

- **Analysis of variance (ANOVA) table**

	df	SS	MSS
Regression	k=1	RSS	MSR=RSS/k
Error	n-2 (n-k-1)	SSE	MSE=SSE/(n-2)
Total	n-1	SST	-

- **Standard error of estimate:**

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

- **Coefficient of determination (R^2)**

$$R^2 = \frac{RSS}{SST} = 1 - \frac{SSE}{SST}$$

$$= \frac{explained\ variation}{total\ variation} = 1 - \frac{unexplained\ variation}{total\ variation}$$

Measure Fitness-F-test

- F test assesses ***the effectiveness of the model as a whole*** in explaining the dependent variable.
- **Define hypothesis:**
  - H_0: $b_1 = b_2 = b_3 = \ldots = b_k = 0$
  - H_a: at least one $b_j \neq 0$ (for j = 1, 2, …, k)
- **F-statistic**$= F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/(n-k-1)}$
- **Critical value (查表)：** F_α (k, n-k-1) "**one-tailed**" F-test; alpha=5%
- **Decision rule**
  - Reject H_0 : if F-statistic > F_α (k, n-k-1)

Critical value: F-Table

Appendix C Table of the F-Distribution

**** Critical values for right-hand tail area equal to 0.05***

Numerator : df_1 and Denominator : df_2

	Df1 : 1	2	3	4	5	…	…	25	30	40	60	120	∞
Df2 : 1	161	200	216	225	230	…	…	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	…	…	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	…	…	8.63	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	…	…	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	…	…	4.52	4.50	4.46	4.43	4.40	4.37
…	…	..	…	…	…	…	…	…	…	…	…	…	…
25	4.24	3.39	2.99	2.76	2.60	…	…	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	…	…	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	…	…	1.78	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	…	…	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	…	…	1.60	1.55	1.50	1.43	1.35	1.25
Infinity	3.84	3.00	2.60	2.37	2.21	…	…	1.51	1.46	1.39	1.32	1.22	1.00

Measure Fitness-SEE

- **Standard Error of Estimate (SEE)**
  - SEE is the standard deviation of the error term. (i.e., the degree of variability of the actual Y-values relative to the estimated Y-values).
  - $SEE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(\varepsilon_i - 0)^2}{n-2}}$
  - SEE measures how well a given linear regression model captures the relationship between the dependent and independent variable.
    - SEE is low if the regression is very strong;
    - SEE is high if the relationship is weak.
  - In ANOVA table, $\boldsymbol{SEE = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}}$

Measure Fitness-R^2

- **Coefficient of determination (R^2) measures** the fraction of the total variation in the dependent variable that is explained by the independent variable.
  - $\mathbf{R^2 = \frac{RSS}{TSS} = 1 - \frac{SSE}{TSS}}$
    - R^2 of 0.8250 means the independent variable explains approximately 82.5 percent of the variation in the dependent variable.
  - $0 \leq R^2 \leq 1$
  - The higher R^2 , the better fitness.

Measure Fitness-Multiple R

- **Multiple R**: the correlation between the actual values and the forecast values of Y.
  - Multiple R=$\sqrt{R^2}$>0.
  - In simple linear regression,Multiple R=$|r_{x,y}|$
    - If b_1>0, Multiple R=$r_{x,y}$;
    - If b_1 <0, Multiple R=-$r_{x,y}$;
- **Correlation vs R^2**
  - Correlation coefficient indicates the sign of the relationship between two variables, whereas R^2(or multiple R) does not.
  - R^2(or multiple R) can apply to multiple regression and implies an explanatory power, while the correlation coefficient only applies to two variables and does not imply explanatory power.

Example

- An analyst ran a regression and got the following result:

	Coefficient	t-statistic	p-value
Intercept	-0.5	-0.91	0.18
Slope	2	20.00	<0.001

ANOVA Table	df	SS	MSS	F
Regression	1	8000	?	400
Error	100	2000	20	
Total	101	?	-	

  - Fill in the blanks
  - What is the standard error of estimate?
  - What is the result of the slope coefficient significance test?
  - What is the result of the sample correlation?
  - What is the 95% confidence interval of the slope coefficient?

Example

- Suppose you run a cross-sectional regression for 100 companies, where the dependent variable is the annual return on stock and the independent variable is the lagged percentage of institutional ownership (INST). The results of this simple linear regression estimation are shown in Exhibit 23. Evaluate the model by answering the questions below.

ANOVA Table for Annual Stock Return Regressed on Institutional Ownership

Source	Sum of Squares	Degrees of Freedom	Mean Square
Regression	576.1485	1	576.1485
Error	1,873.5615	98	19.1180
Total	2,449.7100		

Example

- **1** What is the coefficient of determination for this regression model?
- **2** What is the standard error of the estimate for this regression model?
- **3** At a 5% level of significance, do we reject the null hypothesis of the slope coefficient equal to zero if the critical *F*-value is 3.938?
- **4** Based on your answers to the preceding questions, evaluate this simple linear regression model.

Example

- **Correct Answer:**
  - **1** The coefficient of determination is sum of squares regression/sum of squares total: 576.148 ÷ 2,449.71 = 0.2352, or 23.52%.
  - **2** The standard error of the estimate is the square root of the mean square error: 19.1180 = 4.3724 .
  - **3** Using a six-step process for testing hypotheses, we get the following:

Step 1	State the hypotheses.	H_0: $b_1 = 0$ versus H_a: $b_1 \neq 0$
Step 2	Identify the appropriate test statistic.	$F = \frac{\text{MSR}}{\text{MSE}}$ with 1 and 98 degrees of freedom.
Step 3	Specify the level of significance.	$\alpha = 5\%$ (one tail, right side).
Step 4	State the decision rule.	Critical F-value = 3.938. Reject the null hypothesis if the calculated F-statistic is greater than 3.938.

Example

- **Correct Answer:**
  - **3**

Step 5	Calculate the test statistic.	$F = \frac{576.1485}{19.1180} = 30.1364$
Step 6	Make a decision.	Reject the null hypothesis because the calculated *F*-statistic is greater than the critical *F*-value. There is sufficient evidence to indicate that the slope coefficient is different from 0.0.

  - **4** The coefficient of determination indicates that variation in the independent variable explains 23.52% of the variation in the dependent variable. Also, the *F*-statistic test confirms that the model's slope coefficient is different from 0 at the 5% level of significance. In sum, the model seems to fit the data reasonably well.

Summary

Simple Linear Regression

Hypothesis Test

Calculate the test statistics and interpret the results of testing a single parameter or the whole model

Calculate and explain standard error of estimate and coefficient of determination

Estimate of Y

- 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
- Describe different functional forms of simple linear regressions

Estimate of the Dependent Variable

- **Two sources of uncertainty** when using the regression model and the estimated parameters to make a prediction.
  - The error term itself contains uncertainty.
  - Uncertainty in the estimated parameters.
- **Point estimate**
  - $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$
- **Confidence interval estimate**
  - $\hat{Y} \pm (t_c \times S_f)$
    - t_c = the critical t-value with df=n−2
    - S_f = the standard error of the forecast

$$S_f = SEE \times \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)S_X^2}} = SEE \times \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

Summary

Simple Linear Regression

Estimate of Y

Calculate the point estimate and the confidence interval of estimated Y

Forms of Simple Linear Regression

- Describe a simple linear regression model and the roles of the dependent and independent variables in the model

Forms of simple linear regression

- **Log-Lin Model:** the dependent variable is logarithmic but the independent variable is linear;
  - $\ln Y = b_0 + b_1 X$
  - The slope coefficient in this model is the relative change in the dependent variable for an absolute change in the independent variable.
- **Lin-log model:** the dependent variable is linear but the independent variable is logarithmic; and
  - $\mathrm{Y} = b_0 + b_1 \ln X$
  - The slope coefficient in this regression model provides the absolute change in the dependent variable for a relative change in the independent variable.
- **Log-log model:** where both the dependent and independent variables are in logarithmic form.
  - $\ln Y = b_0 + b_1 \ln X$
  - This model is useful in calculating elasticities because the slope coefficient is the relative change in the dependent variable for a relative change in the independent variable.

Example

- An analyst is investigating the relationship between the annual growth in consumer spending (CONS) in a country and the annual growth in the country's GDP (GGDP). The analyst estimates the following two models:

	Model 1	Model 2
	$GGDP_i = b_0 + b_1 CONS_i + \varepsilon_i$	$GGDP_i = b_0 + b_1 \ln(CONS_i) + \varepsilon_i$
Intercept	1.040	1.006
Slope	0.669	1.994
R^2	0.788	0.867
Standard error of the estimate	0.404	0.320
F-statistic	141.558	247.040

Example

- Questions:
  - **1** Identify the functional form used in these models.
  - **2** Explain which model has better goodness-of-fit with the sample data.

- **Correct Answer:**
  - **1** Model 1 is the simple linear regression with no variable transformation, whereas Model 2 is a Lin-log model with the natural log of the variable CONS as the independent variable.
  - **2** The Lin-log model, Model 2, fits the data better. Since the dependent variable is the same for the two models, we can compare the fit of the models using either the relative measures (R^2 or F-statistic) or the absolute measure of fit, the standard error of the estimate. The standard error of the estimate is lower for Model 2, whereas the R2 and F-statistic are higher for Model 2 compared with Model 1.

Summary

Simple Linear Regression

Forms of Simple Linear Regression

Identify different forms of simple linear regression

Summary

Module: Simple Linear Regression

Basics of Simple Linear Regression

Estimate of regression coefficients

Hypothesis Test

Estimate of Y

Forms of Simple Linear Regression

Module

Introduction to Big Data Techniques

1. How is fintech used in quantitative Investment analysis
2. Advanced Analytical Tools: Artificial Intelligence and Machine Learning
3. Tackling Big Data with Data Science

What is Fintech

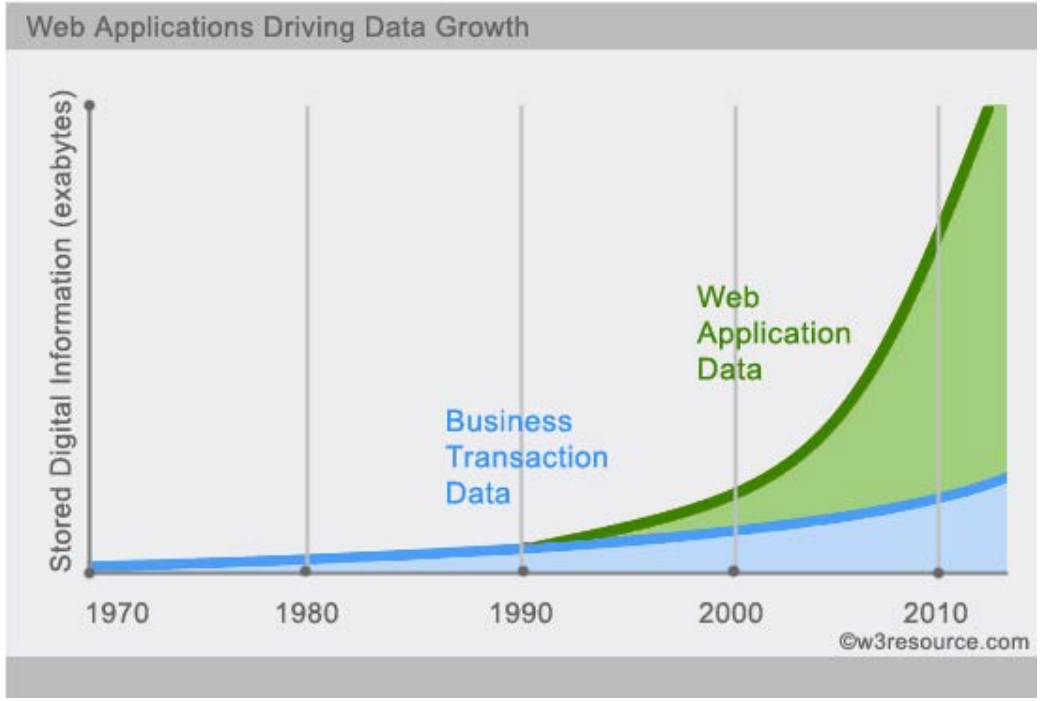
- Areas of Fintech Development
- Analysis of Large Datasets

Areas of Fintech Development

- **Areas of fintech development**
  - Analysis of large datasets
  - Analytical techniques
  - Automated trading
  - Automated advice
  - Financial record keeping

Analysis of Large Datasets

- **Analysis of large datasets**
  - In addition to growing amounts of traditional data, massive amounts of alternative data generated from non-traditional data sources
    - ✓ **Traditional data source**: security prices, corporate financial statements, and economic indicators
    - ✓ **Non-traditional data source**: social media sensor networks



Areas of Fintech Development

- **Analytical tools**
  - **Artificial Intelligence (AI)** – computer systems capable of performing tasks that previously required human intelligence.
- **Automated trading**
  - Computer algorithms or automated trading applications may provide a number of benefits to investors:
    - ✓ i.e. more efficient trading, lower transaction costs, anonymity, and greater access to market liquidity.
- **Automated advice**
  - Robo-advisers or automated personal wealth management services.
  - To provide investment services to retail investors at lower cost.
- **Financial record keeping**
  - New technology, such as Distributed Ledger Technology (DLT), may provide secure ways to track ownership of financial assets on a peer-to-peer (P2P) basis, such as Bitcoin.

Example

Areas of Fintech Development

- A correct description of Fintech is that it:

  A. is driven by rapid growth in data and related technological advances.

  B. increases the need for intermediaries.

  C. is at its most advanced state using systems that follow specified rules and instructions.

- Solution: A.

  Drivers of fintech include extremely rapid growth in data (including their quantity, types, sources, and quality) and technological advances enabling the capture and extraction of information from it.

Summary

Introduction to Big Data Techniques

What is Fintech

Areas of Fintech Development

Analysis of Large Datasets

Big Data

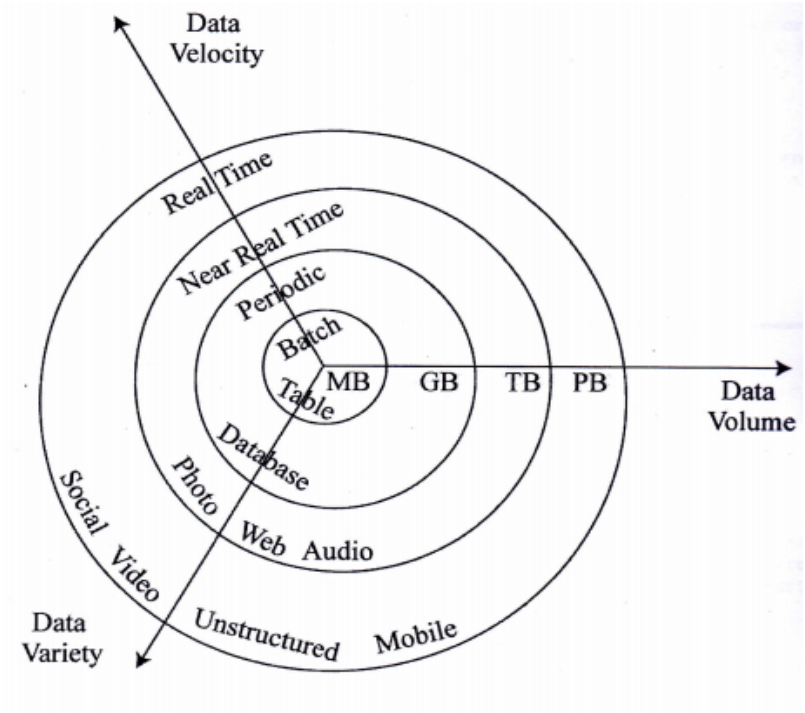
- Characteristics of Big Data
- Type of Big Data
- Main Sources of Alternative Data
- Big Data Challenges

Characteristics of Big Data

- **Definition**
  - The term **Big Data** refers to **the vast amount** of data being generated by industry, governments, individuals, and electronic devices, including data generated from **traditional sources** as well as **non-traditional data types** (also known as **alternative data**)

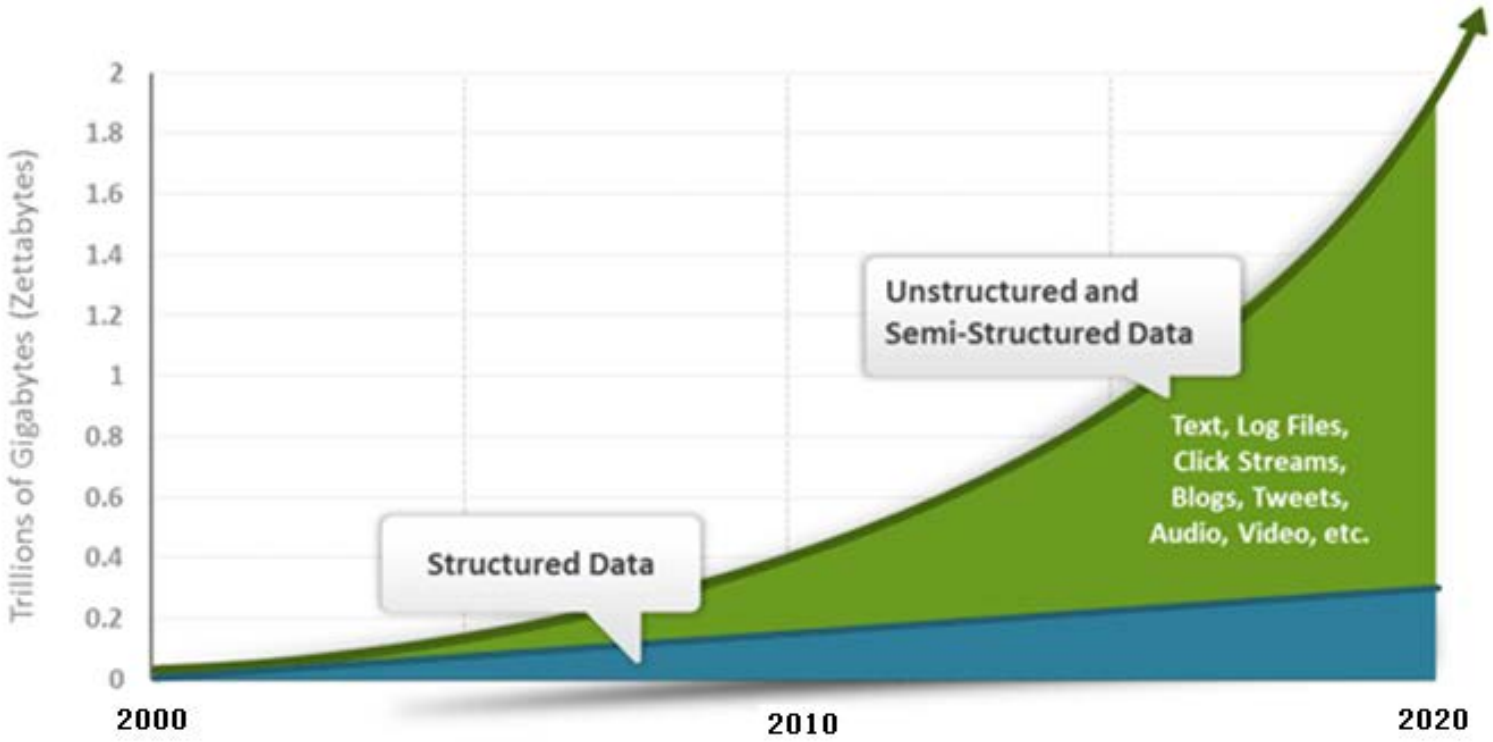
➢ **The characteristics of Big Data**

- Volume (very large)
- Velocity (real-time or near-real-time)
- Variety (mainly unstructured)



Type of Big Data

- **Structured, semi-structured and unstructured data**
  - Structured data : SQL tables or CSV files
  - Semi-structured data : HTML code
  - Unstructured data : video message, blogs, WeChat messages



Main Sources of Alternative Data

- **Three main sources of alternative data**
  - Individuals.
  - Business processes: Including direct sales information, such as credit card data, as well as corporate exhaust.
  - Sensors: Sensor data are collected from such devices as smart phones, cameras, RFID chips, and satellites that are usually connected to computers via wireless networks.

Individuals	Business Processes	Sensors
Social media	Transaction data	Satellites
News, reviews	Corporate data	Geolocation
Web searches, personal data		**Internet of Things**
		Other sensors

Big Data Challenges

- **Big Data challenges**
  - Big Data poses several challenges when it is used in investment analysis, including **the quality, volume, and appropriateness of the data.**
  - The data must be sourced, cleansed, and organized before analysis can occur. This process can be **extremely difficult** with alternative data owing to the unstructured characteristics of the data involved.

Example

Characteristics of Big Data

- A characteristic of Big Data is that:

  A. One of its traditional sources is business processes.

  B. It involves formats with diverse types of structures.

  C. Real-time communication of it is uncommon due to vast content.

- Solution: B.

  Big Data is collected from many different sources and is a variety of formats, including structured data (e.g., SQL tables or CSV files), semi-structured data (e.g., HTML code), and unstructured data (e.g., video messages).

Summary

Introduction to Big Data Techniques

Big Data

Characteristics of Big Data

Type of Big Data

Main Sources of Alternative Data

Big Data Challenges

Artificial Intelligence and Machine Learning

- Advanced Analytical Tools
- Types of Machine Learning

Advanced Analytical Tools

- **Artificial intelligence**
  - Artificial intelligence computer systems are capable of performing tasks that have traditionally required human intelligence. This is often accomplished through the use of *"if then" rules.*
- **Machine learning (ML)**
  - Machine learning (ML) is a technology that has grown out of the wider AI field.
  - ML algorithms are computer programs that are able to *"learn"* how to complete tasks, improving their performance over time with experience.

Advanced Analytical Tools

- **How machine learning works?**
  - Dataset can be split into a **training dataset and validation dataset** (evaluation dataset)
    - ✓ The training dataset allows the algorithm to identify relationships between inputs and outputs based on historical patterns in the data.
    - ✓ These relationships are then tested on the validation dataset.
  - ML still **required human judgement** in understanding data and choosing the right analytic techniques.
  - Errors may arise from *overfitting* and ***underfitted.***
    - ✓ Overfitting: make too much use of the data.
    - ✓ Underfitted: make too little use of the data.
  - ✓ In addition, ML techniques can appear to be opaque **or "black box"** approaches, which arrive at outcomes that **may not be entirely understood or explainable.**

Types of Machine Learning

- **Types of machine learning**
  - **Supervised learning**
    - ✓ Computers learn to model relationships based on **labeled training data**.
    - ✓ Trying to group companies into peer groups based on their industries.
  - **Unsupervised learning**
    - ✓ Computers are not given labeled data but instead are given only data from which the algorithm seeks to describe the data and their structure.
    - ✓ Trying to group companies into peer groups based on their characteristics rather than using standard sector or other acknowledged criteria.
    - ✓ i.e. identify whether it is money laundering, spam mail classification.

Advanced Analytical Tools

- In the use of machine learning (ML):

  A. Some techniques are termed “black box” due to data biases.

  B. Human judgment is not needed because algorithms continuously learn from data.

  C. Training data can be learned too precisely, resulting in inaccurate predictions when used with different datasets.

- Solution: C.

  Overfitting occurs when the ML model learns the input and target dataset too precisely. In this case, the model has been “over trained” on the data and is treating noise in the data as true parameters. An ML model that has been overfitted is not able to accurately predict outcomes using a different dataset and may be too complex.

Summary

Introduction to Big Data Techniques

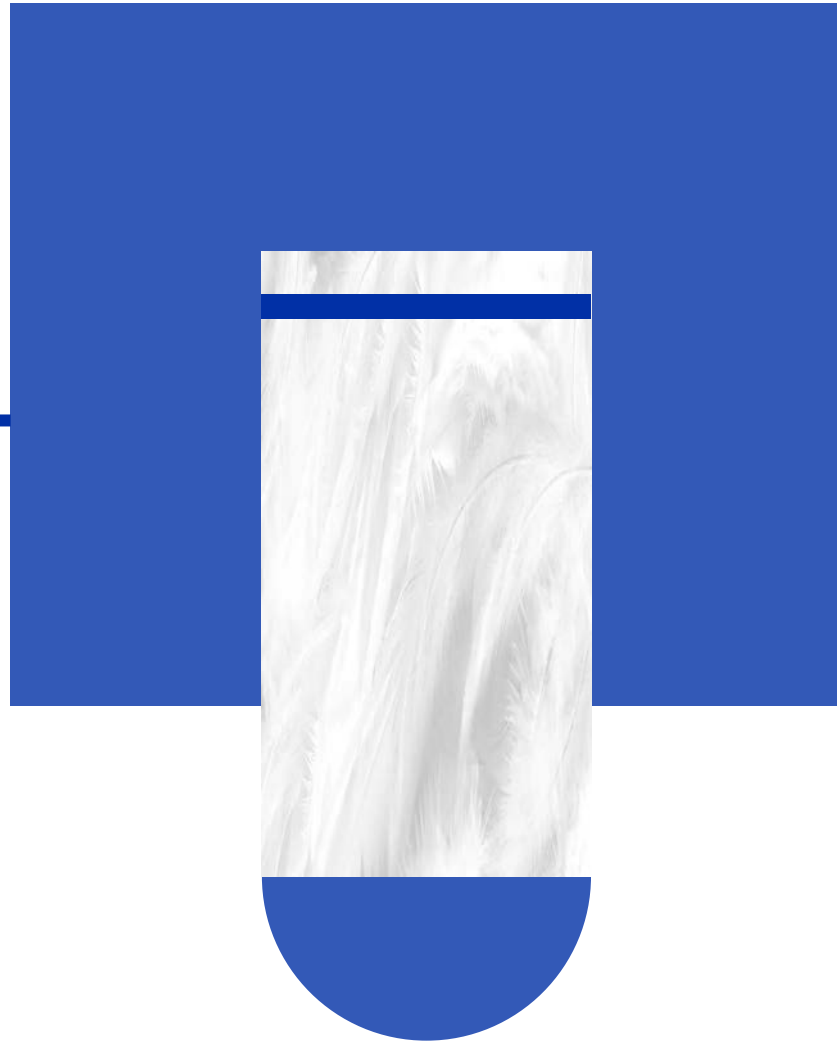
Artificial Intelligence and Machine Learning

Advanced Analytical Tools

Types of Machine Learning

Tackling big data with data science

- Describe applications of Big Data and Data Science to investment management



Tackling big data with data science

- **Data science**
  - is an interdisciplinary field that harnesses advances in computer science (including ML), statistics, and other disciplines for the purpose of extracting information from Big Data (or data in general).
  - Companies rely on the expertise of data scientists/analysts to extract information and insights from Big Data for a wide variety of business and investment purposes.
- **Data processing methods**
  - to help determine the best data management technique needed for Big Data analysis
  - Includes: capture, curation, storage, search, and transfer.

Various data processing methods

- **Capture**
  - Data capture refers to how the data are collected and transformed into a format that can be used by the analytical process.
    - ✓ Low-latency systems (systems that operate on networks that communicate high volumes of data with minimal delay (latency)) are essential for automated trading applications that make decisions based on real-time prices and market events.
    - ✓ In contrast, high-latency systems do not require access to real-time data and calculations.
- **Curation**
  - Data curation refers to the process of ensuring data quality and accuracy through a data cleaning exercise. This process consists of reviewing all data to detect and uncover data errors (bad or inaccurate data) and making adjustments for missing data when appropriate.

Various data processing methods

- **Storage**
  - Data storage refers to how the data will be recorded, archived, and accessed and the underlying database design. An important consideration for data storage is whether the data are structured or unstructured and whether analytical needs require low-latency solutions.
- **Search**
  - Search refers to how to query data. Big Data has created the need for advanced applications capable of examining and reviewing large quantities of data to locate requested data content.
- **Transfer**
  - Transfer refers to how the data will move from the underlying data source or storage location to the underlying analytical tool. This could be through a direct data feed, such as a stock exchange's price feed.

Data Visualization

- As an important tool for understanding Big Data, **visualization** refers to how the data will be formatted, displayed, and summarized in graphical form. Traditional structured data can be visualized using tables, charts, and trends, whereas non-traditional unstructured data require new techniques of data visualization.
- **A word cloud (tag cloud) is used for unstructured, textual data.**
  - A valuable Big Data visualization technique.
  - Words are sized and displayed on the basis of the frequency of the word in the data file.
  - Quickly perceive the most frequent terms.
  - Colors can add an additional dimension.
    - Green:
      - "profit" typically indicates positive sentiment;
    - Red:
      - "loss" typically indicates negative sentiment.

• Text Analytics and Natural Language Processing •

- **Text analytics**
  - Computer programs that analyze and derive meaning typically from large, unstructured text- or voice-based datasets, which include
    - ✓ Company filings, written reports, quarterly earnings calls, social media, email, internet postings, and surveys.
- **An important application of text analytics is natural language processing (NLP).**
  - A computer programs to analyze and interpret human language.
  - Applications include *translation, speech recognition, text mining, sentiment analysis, and topic analysis.*
  - Models using NLP analysis may incorporate **non-traditional information** to evaluate what people are saying such as their preferences, opinions, likes, or dislikes – in an attempt to **identify trends** and short-term indicators about a company, a stock, or an economic event that might have a bearing on future performance.

Summary

Introduction to Big Data Techniques

Tackling big data with data science

Summary

Module: Introduction to Big Data Techniques

How is fintech used in quantitative Investment analysis

Advanced Analytical Tools: Artificial Intelligence and Machine Learning

Tackling Big Data with Data Science

问题反馈

- 如果您认为金程**课程讲义/题库/视频**或其他资料中**存在错误，欢迎您告诉我们**，所有提交的内容我们会在最快时间内核查并给与答复。
- **如何告诉我们？**
  - 将您发现的问题通过扫描右侧二维码告知我们，具体的内容包含:
    - ✓ 您的姓名或网校账号
    - ✓ 所在班级
    - ✓ 问题所在科目(若未知科目，请提供章节、知识点和页码）
    - ✓ 您对问题的详细描述和您的见解
- **非常感谢您对金程教育的支持，您的每一次反馈都是我们成长的动力。**

心有猛虎，细嗅蔷薇。

In me the tiger sniffs the rose.