Quantitative Methods

CFA一级培训项目

讲师：王慧琳

学位证书

- 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持证人
- 通过证券从业资格考试

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。

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。

王慧琳
6年授课，3000+授课课时

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

- **强化班知识点说明和使用指南**

序号	课件元名称（知识点）	必考	高频	低频
6	Annuity	0	1	0
10	Measures of Central Tendency	1	0	0
11	Dispersion	1	0	0
14	Expected value & variance	0	0	1
23	Central limit theory	0	1	0
34	Forms of Simple Linear Regression	0	0	1

- 必考知识点指的是近10年考试中考试频率大于等于75%的考点，在强化班中重点讲解，必须掌握；
- 高频知识点指的是近10年考试中考试频率介于25%到75%的考点，在强化班中重点讲解，必须掌握；
- 低频知识点指的是近10年考试中考试频率小于25%的考点，在基础班中重点讲解，学员可以根据自己的掌握情况在基础班中巩固学习；
- 本学科知识点合计37个，其中必考知识点9个，高频知识点16个，低频知识点12个，掌握必考和高频考点覆盖了近10年90.52%的题目。

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

Framework

Module
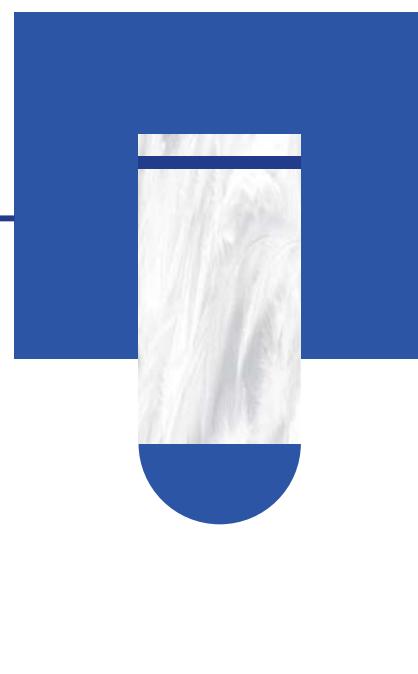
Rates and Returns

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

Annualized return

☐ EAR

☐ Non-annual Compounding

☐ Continuously Compounded Rates of Return

EAR

- **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.

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:

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

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 <u>holding period return</u> 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 <u>price relative</u>, $\frac{S_{t+1}}{S_t}=1+HPR_{t,t+1}=e^{r_{t,t+1}}$;

    - $r_{t,t+1}=Ln(1+HPR_{t,t+1})=Ln\left(\frac{S_{t+1}}{S_t}\right)$

  - A key assumption in many investment applications is that returns are **independently and identically distributed (i.i.d.).**

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

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

# Summary

Rates and Returns

Annualized return

EAR
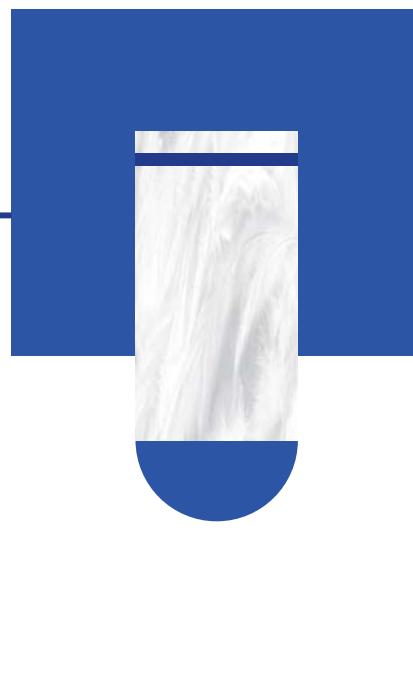
Non-annual Compounding

Continuously Compounded Rates of Return

Money-weighted and time-weighted return

- ☐ Money-weighted rates of return
- ☐ Time-weighted rates of return

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.

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.

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.

Summary

Rates and Returns

Money-weighted and time-weighted return

Money-weighted rates of return

Time-weighted rates of return

Summary

Module: Rates and Returns

Money-weighted and time-weighted return

Annualized return

Module
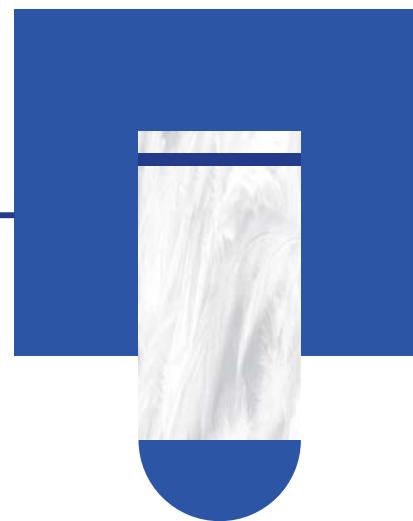
The Time Value of Money in Finance

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

Annuity

☐ Type of annuity

Annuity

- **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)
  - **Perpetuity:** A perpetuity is a set of level never-ending sequential cash flows, with the first cash flow occurring one period from now.
    - Calculation: $PV = \dfrac{A}{1+r} + \dfrac{A}{(1+r)^2} + \dfrac{A}{(1+r)^3} + \cdots = \dfrac{A}{r}$

Summary
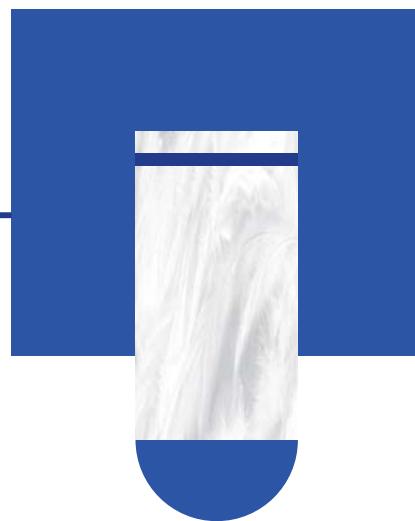
The Time Value of Money in Finance

Annuity

Type of annuity

Cash flow additivity

- ☐ Implied forward rates
- ☐ Forward exchange rates
- ☐ Option pricing

---

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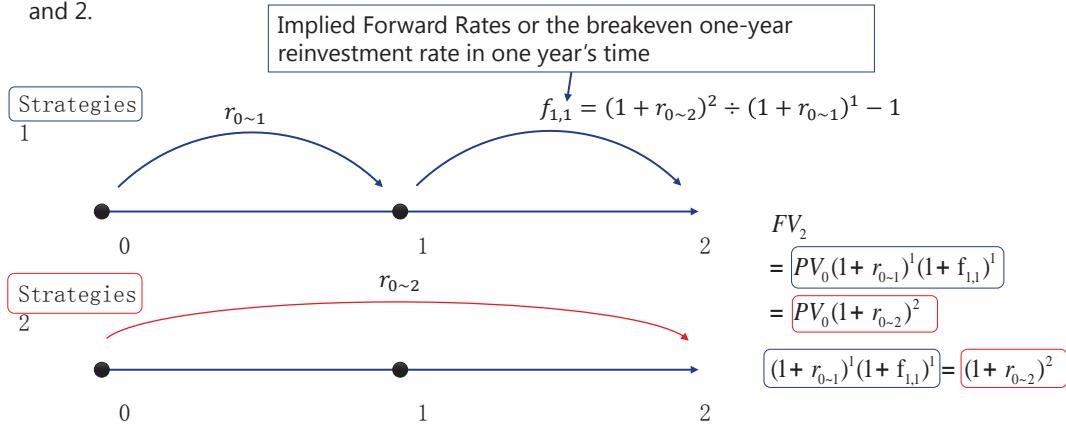
- **Cash flow additivity principle**
  - Under cash flow additivity, the present value of any future cash flow stream indexed at the same point equals the sum of the present values of the cash flows. No possibility exists to earn a riskless profit in the absence of transaction costs.
  - Three economic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

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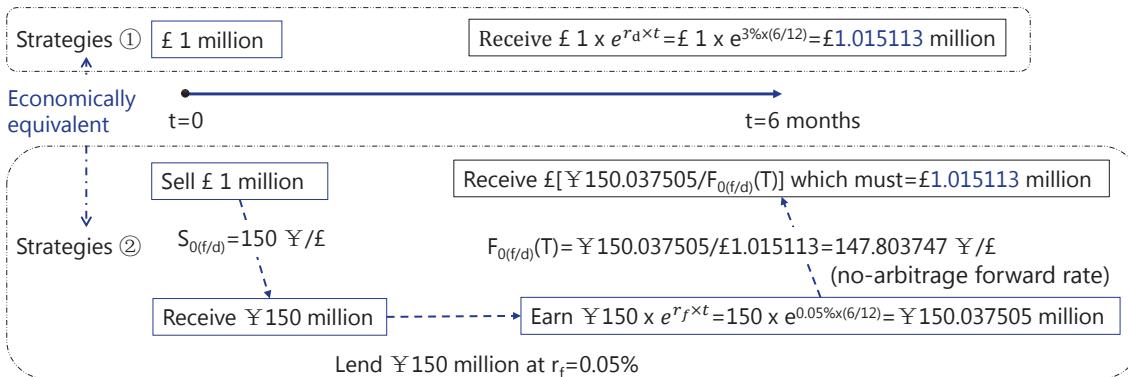
- **Implied Forward Rates Using Cash Flow Additivity**
  - Under the cash flow additivity principle, a risk-neutral investor would be indifferent between strategies 1 and 2.

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

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

Strategies 1

$r_{0\sim1}$

| 0 | 1 | 2 |

Strategies 2

$r_{0\sim2}$

| 0 | 1 | 2 |

FV_2

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

$$= \boxed{PV_0(1 + r_{0\sim2})^2}$$

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

Forward Exchange Rates

- **No-arbitrage forward exchange 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).
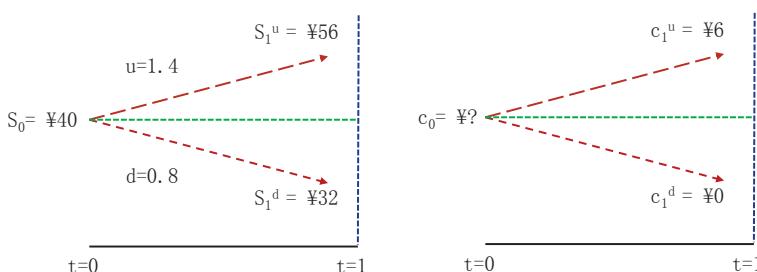
Strategies ① £ 1 million   Receive £ 1 x $e^{r_d \times t}$ = £ 1 x $e^{3\% \times (6/12)}$ = £1.015113 million

Economically equivalent

t=0   t=6 months

Strategies ②   Sell £ 1 million   Receive £[¥150.037505/$F_{0(f/d)}$(T)] which must=£1.015113 million

$S_{0(f/d)}$=150 ¥/£ $F_{0(f/d)}$(T)=¥150.037505/£1.015113=147.803747 ¥/£
(no-arbitrage forward rate)

Receive ¥150 million ------- Earn ¥150 x $e^{r_f \times t}$=150 x $e^{0.05\% \times (6/12)}$= ¥150.037505 million

Lend ¥150 million at r_f=0.05%

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).

S_1^u = ¥56 c_1^u = ¥6

u=1. 4

S_0= ¥40 c_0= ¥?

d=0. 8

S_1^d = ¥32 c_1^d = ¥0

t=0   t=1   t=0   t=1
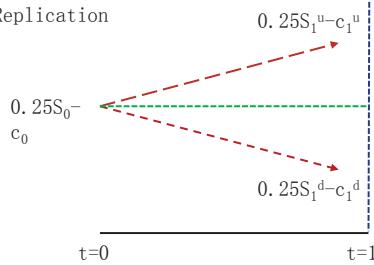
One-Period Binomial Tree for the stock's Price   One-Period Binomial Tree for the Contract's Price

Option Pricing

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

Call Option Replication

$0.25S_1^u - c_1^u$

$0.25S_0 - c_0$

$0.25S_1^d - c_1^d$

$t=0$ $t=1$

$+S$ $+ -c$ $\rightarrow$ $\triangle$portfolio's value $_{t=1}=V_1=0$

$+hS$ $+ -c = 0$

$+ h\triangle S$ $+ - \triangle c = 0$

$h = \triangle c/\triangle S = (6-0)/(56-32) = 0.25 =$ Hedge ratio

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

$V_0 = 0.25S_0 - c_0$

$V_0 = V_1$

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

$c_0 = 0.25 \times 40 - 8/(1+R_f)^1$

Summary

The Time Value of Money in Finance

Cash flow additivity

Implied forward rates

Forward exchange rates

Option pricing

Summary

Module: The Time Value of Money in Finance

Annuity

**Cash flow additivity**

Module
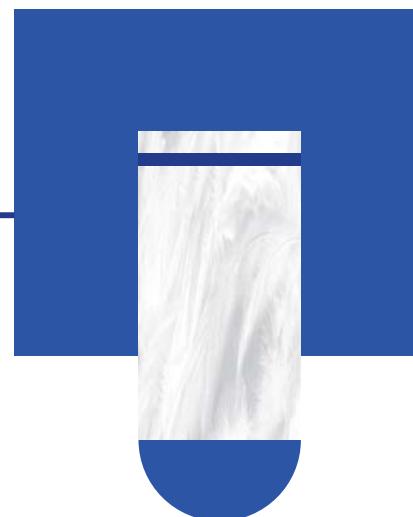
Statistical Measures of Asset Returns

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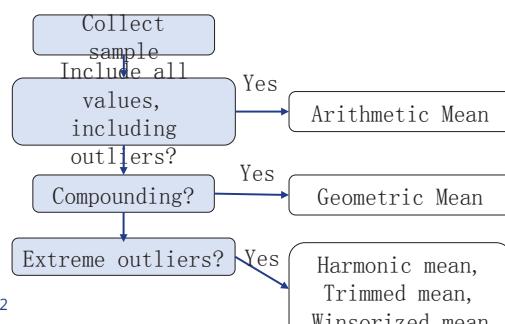
Measures of central tendency and location

□ Measures of Central Tendency
□ Dealing with Outliers

Measures of central tendency

- **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.

Collect sample

Include all values, including outliers? —Yes→ Arithmetic Mean

Compounding? —Yes→ Geometric Mean

Extreme outliers? —Yes→ Harmonic mean, Trimmed mean, Winsorized mean
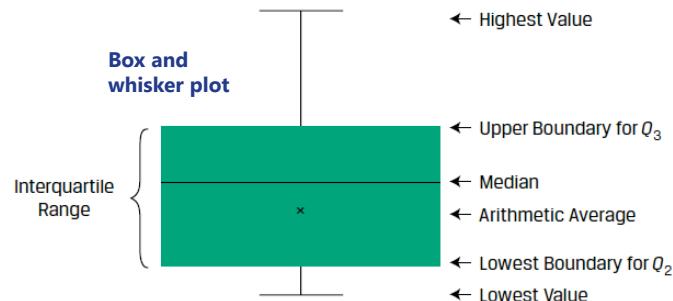
Measures of central tendency

- **Median is** the value of the middle item of a set of items that has been <u>sorted into</u> ascending or descending order.
    - In an <u>odd-numbered</u> sample of n items, the median is the value of the item that occupies the (n + 1)/2 position.
    - In an <u>even-numbered</u> 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.
- **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.

Measure of Location

- **Quartile /Quintile/Decile/Percentile.**
    - The third quintile: there are 60% the observations fall at or below that value.
- **Calculation: Ly = (n+1)y/100.**

Box and whisker plot

← Highest Value

← Upper Boundary for Q_3

Interquartile Range

← Median
← Arithmetic Average

← Lowest Boundary for Q_2
← Lowest Value

Example

Observers: 5 8 11 12 14 16 16 18 19 21 23   **Calculate the third quartile** of the data set

Correct Answer:

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

The third quartiles = 19

Summary

Statistical Measures of Asset Returns
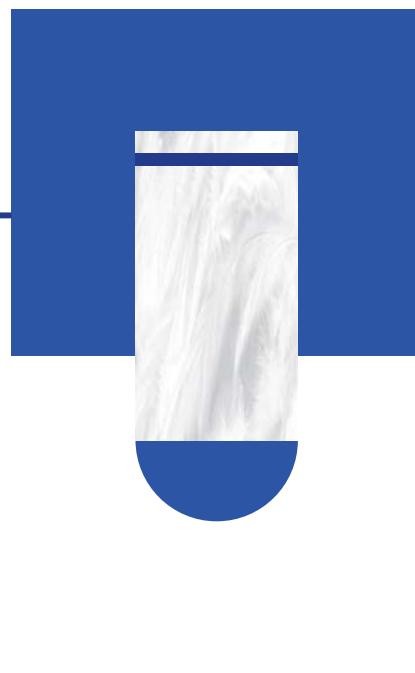
Measures of Central Tendency and Location

Measures of Central Tendency

Dealing with Outliers

Measures of Dispersion

☐ Absolute Dispersion
☐ Relative Dispersion

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)** $= \frac{\sum_{i=1}^{N} |X_i - \overline{X}|}{n}$

  - **Variance (Var) & standard deviation (S.D.)**

    - **For population:** $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ $\qquad \sigma = \sqrt{\sigma^2}$

    - **For sample:** $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ $\qquad s = \sqrt{s^2}$

  - **Downside deviation**

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

- 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.

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

  - Relative dispersion.
  - Scale free.

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

$$\text{Sharpe ratio} = \frac{E(R_P) - R_f}{\sigma_P}$$

Summary

## Statistical Measures of Asset Returns
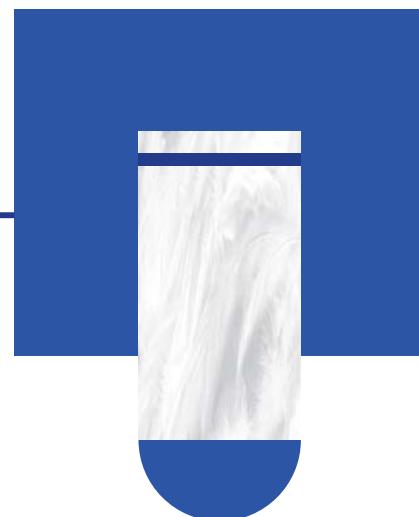
Measures of Dispersion

Absolute Dispersion

Relative Dispersion
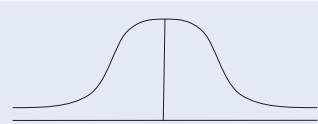
Measures of Shape of a Distribution

- ☐ Measures of Skewness
- ☐ Measures of Kurtosis
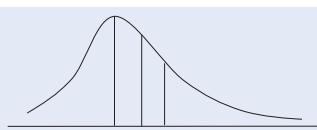
Skewness
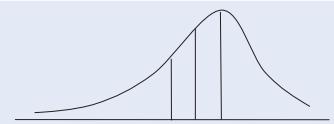
Mean=Median=Mode	Mode<Median<Mean	Mean<Median<Mode
Symmetrical	**Positive (right) skew**	**Negative (left) skew**
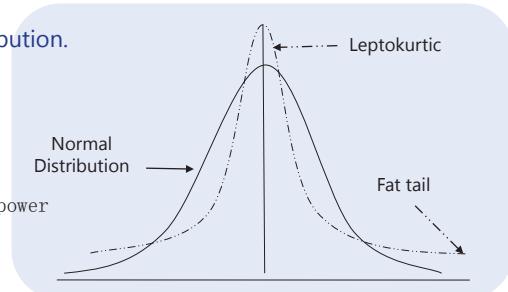
- A distribution that is not symmetrical is termed **skewed.**
  - **Positively skewed**: Mode<median<**mean**, having a long tail on the **right** side.
    - A return distribution with positive skew has frequent small losses and few extreme gains.
  - **Negatively skewed**: Mode>median>**mean**, having 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:**

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

the third power

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

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

Leptokurtic

Normal Distribution

Fat tail

Summary

Statistical Measures of Asset Returns

Measures of Shape of a Distribution

Measures of Skewness

Measures of Kurtosis

Summary

Module: Statistical Measures of Asset Returns

Measures of Central Tendency

Measures of Dispersion

### Measures of Shape of a Distribution

Module
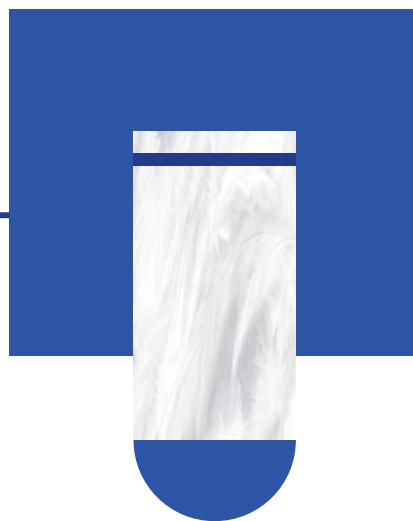
## Probability Trees and Conditional Expectations
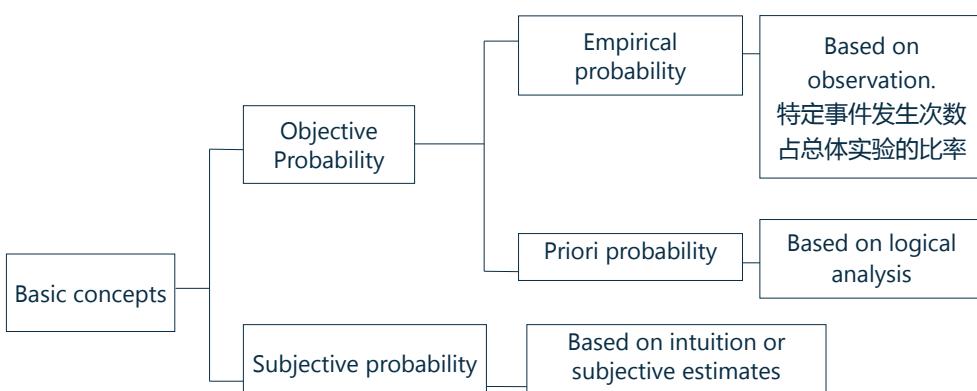
1. Calculation Rules for Probabilities
2. Probability trees and conditional expectations
3. Bayes' Formula and Updating Probability Estimates

Calculation Rules for Probabilities

☐ Calculation Rules for Probabilities

Probability Concepts

```
                              ┌─ Empirical ──── Based on
                              │   probability    observation.
                              │                  特定事件发生次数
              ┌─ Objective ───┤                  占总体实验的比率
              │   Probability  │
              │               └─ Priori ──────── Based on logical
Basic concepts ┤                  probability     analysis
              │
              └─ Subjective ──── Based on intuition or
                  probability      subjective estimates
```
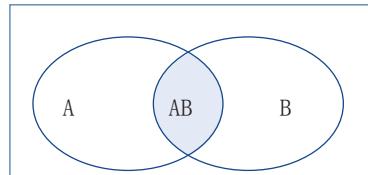
Calculation Rules for Probabilities

- **Multiplication rule**
  - Probability that two events will happen at the same time: Joint probability P(AB)
    - $P(AB)=P(A|B)\times P(B)= P(B|A)\times P(A)$
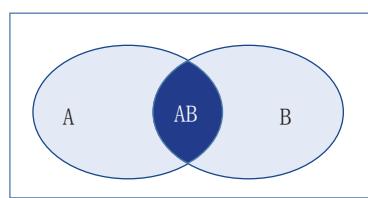
- **Addition rule**
  - The probability that A or B occurs, or both occur
    - $P(A \text{ or } B)=P(A)+P(B)-P(AB)$

Calculation Rules for Probabilities

- **Mutually exclusive events**
  - $P(AB)=P(A|B)=P(B|A)=0$ $P(A \text{ or } B)=P(A)+P(B)$
- **Independent events**

  If exclusive, must NOT independence.

  - The occurrence of A <u>doesn't affect</u> the occurrence of B.
    - $P(A|B)=P(A)$ or $P(B|A)=P(B)$
    - $P(AB)=P(A)\times P(B)$ $P(A \text{ or } B)=P(A)+P(B)- P(AB)$

- **Dependent events**
  - The probability of occurrence of A is related to the occurrence of B.

Expected Value & Variance

- **Expected value and variance of a random variable**
  - $E(X) = \sum x_i * P(x_i) = x_1 * P(x_1) + x_2 * P(x_2) + \cdots + x_n * P(x_n)$

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

  - $\sigma = \sqrt{\sigma^2}$
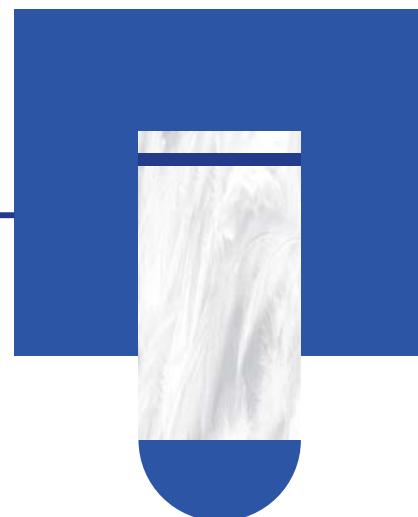
Summary

Probability Trees and Conditional Expectations

Calculation Rules for Probabilities

Bayes' formula and updating probability estimates

☐ Bayes' formula

Bayes' Formula

- P(AB)=P(A|B)×P(B) =P(B|A)×P(A)

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

    - Where P(B) can be solved using **total probability formula**:

        - P(B)=P(B|W_1)×P(W_1)+P(B|W_2)×P(W_2)+...+ P(B|W_n)×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.**

Summary

Probability Trees and Conditional Expectations

Bayes' formula and updating probability estimates

Bayes' Formula

Summary

Module: Probability Trees and Conditional Expectations

Expected Value and Variance

Probability Trees and Conditional Expectations

### Bayes' Formula and Updating Probability Estimates

Module
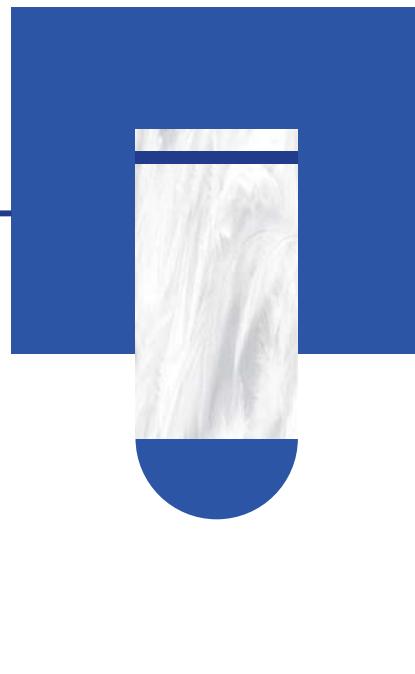
Portfolio Mathematics

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

☐ Portfolio expected return and variance
☐ Covariance & Correlation

Covariance & Correlation

- **Covariance**
  - Covariance is a measure of the co-movement between random variables.
    $$COV(X,Y) = E[(X-E(X))(Y-E(Y))] \qquad Cov(R_i,R_j) = E\left[\left(R_i - ER_i\right)\left(R_j - ER_j\right)\right]$$
  - The covariance of a random variable with itself is its own variance.
    $$COV(X,X) = E[(X-E(X))(X-E(X))] = \sigma^2(X)$$
  - Covariance ranges from negative infinity to positive infinity.
- **Correlation**
  - Correlation measures the co-movement (linear association) between two random variables.
    $$\rho_{XY} = \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}}$$
  - 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.

Covariance & Correlation

- 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 <u>contain information</u> about the two variables' relationship. Otherwise, exclude the outliers.
  - Spurious correlation
    - correlation between two variables that reflects <u>chance relationships</u> in a particular dataset;
    - correlation <u>induced by a calculation</u> that mixes each of two variables with a <u>third variable</u>;
    - correlation between two variables arising <u>not from a direct relation</u> between them but from their relation to a third variable.

- **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).

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

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

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

Summary

Portfolio Mathematics

Portfolio Expected Return and Variance of Return

Covariance & Correlation

Portfolio Risk Measures: Applications of the Normal Distribution

- ☐ Roy's safety-first criterion
- ☐ Managing financial risk tools

Roy's safety-first criterion

- **Shortfall risk: R_L = threshold level return, minimum return required**
  - Minimize $(R_p < R_L)$
- **Roy's safety-first criterion**
  - $[E(R_P) - R_L]/\sigma_P$
- **Maximize S-F-Ratio**
  - Maximize $SFR = \dfrac{E(R_P) - R_L}{\sigma_P}$ <=> Minimize $P(R_p < R_L)$

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).

Summary

Portfolio Mathematics

Portfolio Risk Measures: Applications of the Normal Distribution

Roy's safety-first criterion

Managing financial risk tools

Summary

Module: Portfolio Mathematics

Portfolio Expected Return and Variance of Return

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

❑ Lognormal Distributions
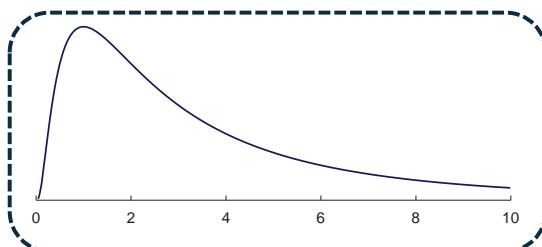
Lognormal Distribution

- **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 <u>is useful for modeling asset prices.</u>

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 <u>very complex</u> and we must <u>assume a parameter distribution in advance</u>.
  - Monte Carlo simulation provides only statistical estimates, <u>not exact results</u>.

Summary

Module: Simulation Methods

Lognormal distribution and continuous compounding

Lognormal distribution

Module
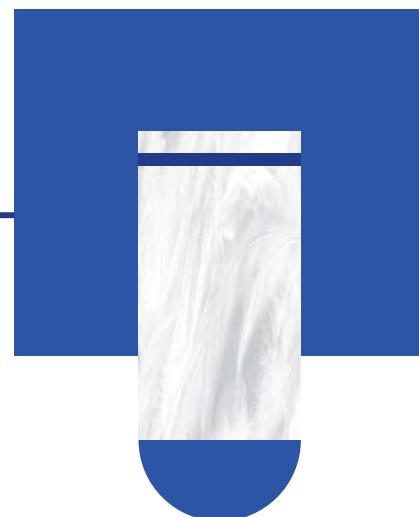
Sampling and Estimation

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

Sampling Methods

☐ Sampling Methods

Sampling Methods

```
                              ┌─ Simple Random Sampling
                              │
                              ├─ Systematic Sampling
          ┌─ Probability ─────┤
          │   Methods         ├─ Stratified Random Sampling
          │                   │
Sampling ─┤                   └─ Cluster Sampling
Methods   │
          │   Non-Probability ┌─ Convenience Sampling
          └─ Methods ─────────┤
                              └─ Judgment Sampling
```
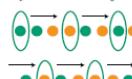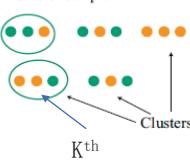

Simple random sample   Stratified sample   Systematic sample   Cluster sample

Subdivision (strata)   Subdivision (strata)

K^{th} Clusters

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.

Summary

Sampling and Estimation

Sampling Methods

Central Limit Theory

- ☐ Central Limit Theory

Central Limit Theory

- **Central Limit Theory**

  - For simple random samples of size n from a population with a mean μ and a finite variance σ² but <u>without known distribution</u>, the <u>sampling distribution of the sample mean</u> approaches <u>N(μ , σ²/n)</u> 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 Theory

Bootstrapping and empirical sampling distributions

- Bootstrap
- Jackknife

Resampling

- **Resampling:** repeatedly draws samples from the <u>original observed data sample</u> for the statistical inference of population parameters.

- **Two types of resampling**

  - **Bootstrapping** uses computer simulation for statistical inference <u>without using an analytical formula such as a z-statistic or t-statistic</u> (model-free resampling or non-parametric resampling)
    - **Bootstrap** usually <u>gives different results</u> because bootstrap resamples are randomly drawn.
      - Bootstrap needs to determine <u>how many repetitions</u> are appropriate.

  - **Jackknife** samples are selected by taking the original observed data sample and <u>leaving out one observation at a time</u> from the set (and not replacing it).
    - **Jackknife** <u>produces similar results for every run</u>.
      - Jackknife usually requires <u>n repetitions</u>. (n=sample size).

Summary

Sampling and Estimation

Bootstrapping and Empirical Sampling Distributions

Summary

Module: Sampling and Estimation

Sampling methods

**Central limit theorem and inference**

Module
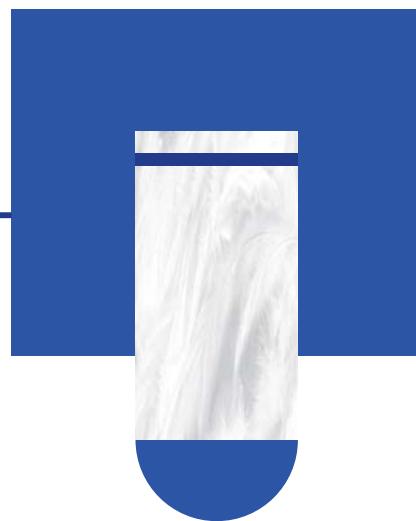
Hypothesis Testing

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

---
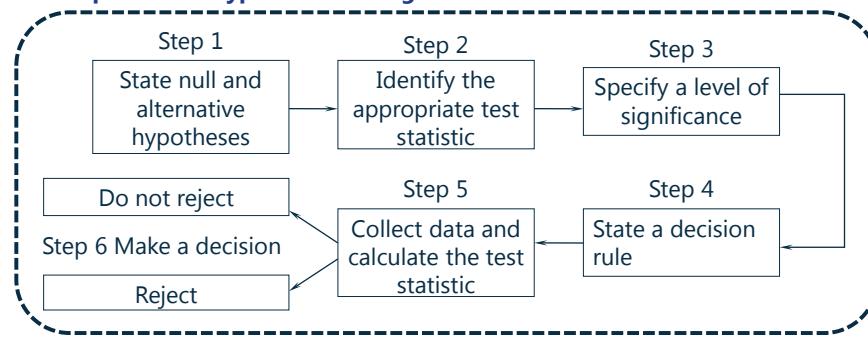
Hypothesis tests for finance

- ☐ Hypothesis testing
- ☐ P-value
- ☐ Type I errors or Type II errors

Hypothesis Testing

- **The process of hypothesis testing**

Step 1	Step 2	Step 3
State null and alternative hypotheses	Identify the appropriate test statistic	Specify a level of significance

Step 6 Make a decision	Step 5	Step 4
Do not reject / Reject	Collect data and calculate the test statistic	State a decision rule

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

Hypothesis Testing

- **Step one: state the hypothesis**

  - Statistical assessment of a statement or idea regarding a <u>population parameter</u>.
  - 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** | $H_0 : \mu = \mu_0$ | $H_a : \mu \neq \mu_0$ |

| **One-tailed** | $H_0 : \mu \leq \mu_0$
 $or, H_0 : \mu \geq \mu_0$ | $H_a : \mu > \mu_0$
 $H_a : \mu < \mu_0$ |

Hypothesis Testing

- **Step two: identify the appropriate test statistic**

  - Test statistic $= \dfrac{Sample\ statistics - Hypothesized\ value}{stanard\ error\ of\ the\ sample\ statistic}$

    - 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**

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

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

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.
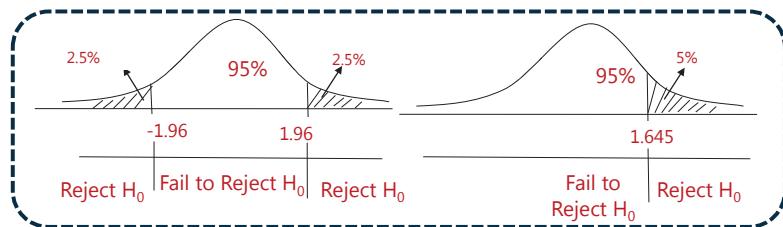
- **Step four: state a decision rule**

  - 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 ******

Hypothesis Testing

- **Relation between Confidence Intervals and Hypothesis Tests**
  - Confidence Interval = [sample statistic ± (critical value) x (standard error)]
  - Center of Interval = sample statistic
  - Length of Interval = 2 x (critical value) x (standard error)

P-value

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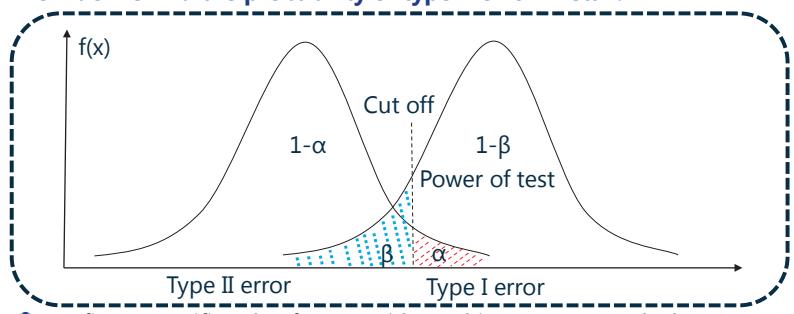
Example
- The p-value for a two-tailed test of sample mean is 1.68%. Which of the following is true?
  - A. We can reject the null with 95% confidence
  - B. We can reject the null with 99% confidence
  - C. the largest probability of rejecting the null hypothesis is 1.68%

- 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 $= P(\text{Type I error}) = P(\text{Reject} \mid H_0 \checkmark)$
- **Type II error: 取伪，fail to reject the null hypothesis when it's actually false**
  - $P(\text{Type II error}) = P(\text{Fail to reject} \mid H_0 \times) = \text{Beta}$
  - **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** $= 1 - P(\text{Type II error}) = P(\text{Reject} \mid H_0 \times)$

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. $P(H_0 \checkmark \mid H_0 X) = P(Ha\ X \mid Ha\checkmark)$
- power of test $= 1 - \text{Beta} = P(Ha\checkmark \mid Ha\checkmark) = 1 - P(H_0 \checkmark \mid H_0 X) = 1 - P(Ha\times \mid Ha\checkmark)$

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.

Summary

Hypothesis Testing

Hypothesis tests for finance
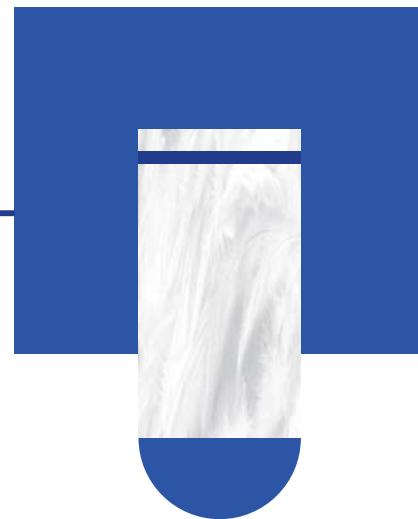
Hypothesis testing

P-value

Type I errors or Type II errors

Tests of return and risk in finance

- Mean hypothesis testing
- Variance hypothesis testing

Summary of Hypothesis Testing

Test type	Assumptions	H_0	Test-statistic	Critical value and degree of freedom
Mean hypothesis testing	Normally distributed population, <u>known</u> population variance	$\mu=0$	$Z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	N(0,1)
	Normally distributed population, <u>unknown</u> population variance	$\mu=0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	t(n-1)
	<u>Independent</u> populations, <u>unknown</u> population variances **assumed equal**	$\mu_1-\mu_2=0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$ $where\ s_p^2\ is\ a\ pooled\ estimator$ $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	t(n$_1$ + n$_2$ - 2)
	Samples <u>not independent</u> (paired comparisons test)	$\mu_d=0$	$t = \dfrac{\bar{d}}{s_{\bar{d}} = s_d/\sqrt{n}}$	t(n-1)

Summary of Hypothesis Testing

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

Summary

Hypothesis Testing

Tests of return and risk in finance

Mean hypothesis testing

Variance hypothesis testing

Parametric versus nonparametric tests

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

- **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.**

Summary

Hypothesis Testing

Parametric versus Nonparametric Tests

Summary

Module: Hypothesis Testing

Hypothesis tests for finance
### Tests of return and risk in finance

Module
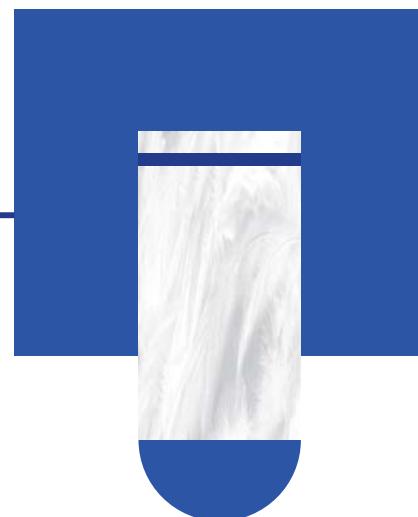
Parametric and Non-Parametric Tests of Independence

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

Tests of independence using contingency table data

☐ Tests of independence

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				Low Risk	High Risk	
Growth	73	26	99		Growth	74%	26%	100%
Value	183	33	216		Value	85%	15%	100%
	256	59	315					

○ Expected Valuei,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				Low Risk	High Risk	
Growth	80.457	18.543	99		Growth	81%	19%	100%
Value	175.543	40.457	216		Value	81%	19%	100%
	256	59	315					

Summary

Module: Hypothesis Testing

### Tests of Independence Using Contingency Table Data

---

Module
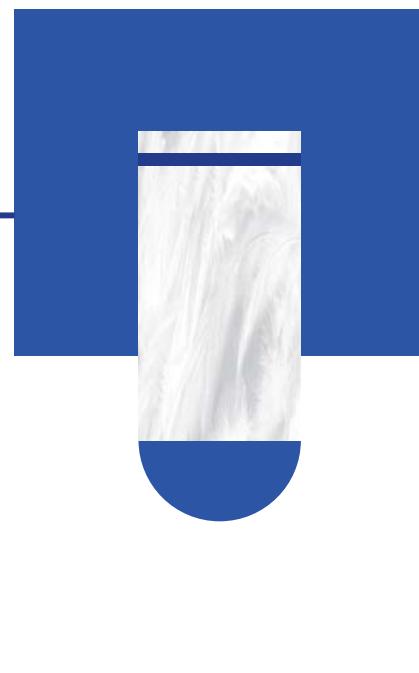
● ● ●

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

- ☐ Interpretation of the parameters
- ☐ Assumptions of linear regression

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.

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.

Summary

Simple Linear Regression
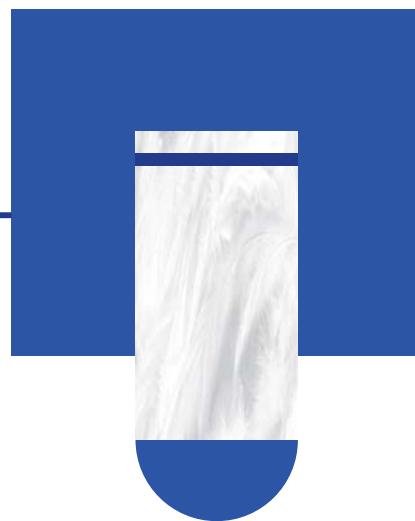
Basics of Simple Linear Regression

Interpretation of the parameters

Assumptions of 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:** $\hat{b}_1 = b_1 \quad \hat{b}_0 = b_0$
  - **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.
- Confidence interval $\hat{b}_1 \pm t_c S_{\hat{b}_1}$

  - $S_{\hat{b}_1}$ 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

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

☐ Test of parameters

☐ Measure of model fitness

Significance test for regression coefficient

- **Significance test for regression coefficient**
  - H_0: $b_1 = 0$
  - Test Statistic= $t = \dfrac{\hat{b}_1 - 0}{S_{\hat{b}_1}} = \dfrac{\hat{b}_1}{S_{\hat{b}_1}}$
  - 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.

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-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).
    - $SEE = \sqrt{\dfrac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}} = \sqrt{\dfrac{\sum_{i=1}^{n}(\varepsilon_i-0)^2}{n-2}}$
    - 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, $SEE = \sqrt{\dfrac{SSE}{n-k-1}} = \sqrt{MSE}$

Measure Fitness-R^2

- **Coefficient of determination (R^2) measures** the fraction of the total variation in the dependent variable that is <u>explained</u> by the independent variable.
    - $R^2 = \dfrac{SSR}{SST} = 1 - \dfrac{SSE}{SST}$
        - 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.

Measure Fitness-ANOVA Table

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

	df	SS	MSS
Regression	k=1	SSR	MSR=SSR/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)**

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

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

Measure Fitness-F-test

- F test assesses ***the effectiveness of the model as a whole*** in explaining the dependent variable.

- **Define hypothesis:**

  - H_0: $b_1 = b_2 = b_3 = ... = b_k = 0$
  - H_a: at least one $b_j \neq 0$ (for j = 1, 2, ..., k)

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

- **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)

Summary

Simple Linear Regression

Hypothesis Test

Measure of Model Fitness

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

Summary

Module: Simple Linear Regression

Basics of Simple Linear Regression

Hypothesis Test

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

## Tackling big data with data science

❑ Data processing methods
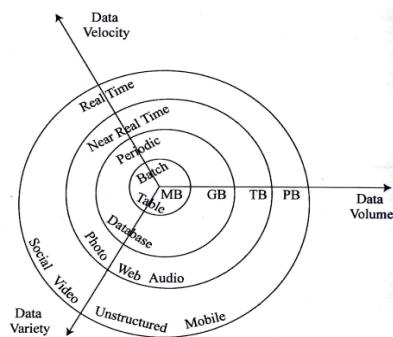
Characteristics of Big Data

- **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)

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))
- **Curation**
  - Data curation refers to the process of ensuring data quality and accuracy through a data cleaning exercise.
- **Storage**
  - Data storage refers to how the data will be recorded, archived, and accessed and the underlying database design.
- **Search**
  - Search refers to how to query data.
- **Transfer**
  - Transfer refers to how the data will move from the underlying data source or storage location to the underlying analytical tool.

Summary

Module: Introduction to Big Data Techniques

Tackling Big Data with Data Science

Data processing methods

Summary

Introduction to Big Data Techniques
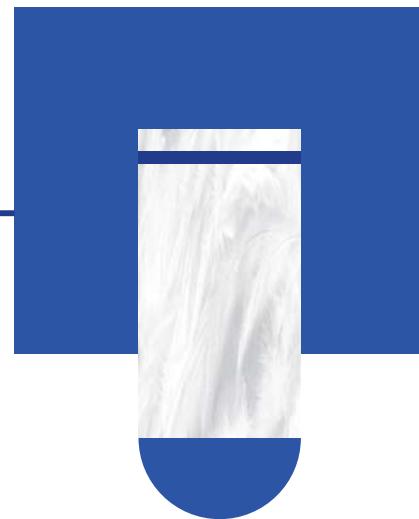
How is fintech used in quantitative Investment analysis

Fintech

Big data

Advanced analytical tools: artificial Intelligence and machine learning

☐ Describe Artificial intelligence, and machine learning

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.

Summary

Introduction to Big Data Techniques

Advanced analytical tools: artificial Intelligence and machine learning

Big Data

Artificial intelligence

Machine learning

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.