王牌陈 讲CFA

高顿教育 GOLDEN EDUCATION

极速串讲

Quantitative Methods

定量分析方法

Level I

王牌陈 讲CFA

陈一磊，CFA/FRM双证持证人，高顿教育资深总监，首席讲师

- **学神级别的开挂人生**
  - 中考、高考、研究生一路以第1名身份保送
  - 本科阶段就读复旦大学财务金融系，GPA全系第一，获管理学学士学位和法学学士学位；研究生阶段就读复旦大学管理学院
  - 以全优成绩通过CFA三个级别考试；一天时间以全优成绩同时通过FRM两个级别考试
- **财经讲师的王牌之路**
  - 全职加入高顿财经前，就职于国有商业银行总行和华尔街投行，同时以兼职身份承担高顿CFA/FRM教学工作
  - 逾12年教龄，CFA/FRM培训界的教父级人物，学员遍布全球

王牌陈讲CFA

小红书

抖音 陈一磊讲CFA

一 利率和投资回报率

1.1 利率的理解

- An **interest rate** is the rate which is charged or paid for the use of money
- Applications of interest rate
  - Required rate of return is the minimum rate of return an investor must receive in order to accept the investment
  - Discounted rate is the rate at which we discount the future amounts to find their value today
  - Opportunity cost is the value that investors forgo by choosing a particular course of action

1.2 利率的组成

- **nominal interest rate = nominal risk free interest rate + risk premiums**
  - nominal risk free interest rate = real risk free interest rate + inflation rate
  - risk premiums = default risk premium + liquidity premium + maturity premium

2.1 持有期间回报率

- **Holding period return** is the return for a single specified period of time
  - $HPR = (P_{end} - P_{begin} + Income_{end}) / P_{begin}$
  - $R_{total} = (1 + HPR_1) \times (1 + HPR_2) \times \cdots \times (1 + HPR_n) - 1$

2.2 平均回报率

- **Arithmetic mean return** focus on average single-period performance
  - $R_{arithmetic} = (R_1 + R_2 + \cdots + R_n) / n$
- **Geometric mean return** focus on average multi-period performance account for **compounding**
  - $R_{geometric} = \sqrt[n]{(1 + R_1)(1 + R_2) \cdots (1 + R_n)} - 1$
- **Harmonic mean return**
  - $R_{harmonic} = \frac{n}{\left(\frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n}\right)}$

- Harmonic Mean ≤ Geometric Mean ≤ Arithmetic Mean
  - The equal sign will only be valid given all the observations are the same
  - The greater variability of the different observation, the more the arithmetic mean will exceed the geometric mean and harmonic mean as well

3. 时间加权回报率和货币加权回报率

- **Time-weighted return (TWR)** is the compound return that $1 initially invested in the portfolio over a stated measurement period
  - $TWR = \sqrt[n]{(1 + HPR_1) \times (1 + HPR_2) \times \ldots \times (1 + HPR_n)} - 1$
- **Money-weighted return (MWR)** is the IRR based on the cash flows related to the investment
  - $CF_0 + CF_1 / (1 + MWR)^1 + \cdots\cdots + CF_n / (1 + MWR)^n = 0$

TWR 【geometric mean】	MWR 【IRR】
HPR → geometric mean → TWR	**cash flow** → IRR → MWR
Periods can be any length between significant CF	Periods must be equal length: use shortest period with no CF
Not affected by cash withdrawals or additions	**Affected** by cash withdrawals or additions
multi-period performance, is used to evaluate fund manager's performance	

MWR assigns more weights to the return of larger cash flows

If more funds to invest at an unfavorable time (favorable time), MWR < TWR (MWR > TWR)

4.1 现值和终值因子

- periodic interest rate = quoted interest rate (r_s) / compounded frequency (m)
- Future value factor $(1 + r_s/m)^{m \times N}$
  - $FV_N = PV \times (1 + r_s/m)^{m \times N}$
- Present value factor $(1 + r_s/m)^{-m \times N}$
  - $PV = FV_N \times (1 + r_s/m)^{-m \times N}$

4.2 年化收益率

- $R_{annual} = (1 + R_{period})^m - 1 = (1 + r_s/m)^m - 1$

4.3 连续复利模式

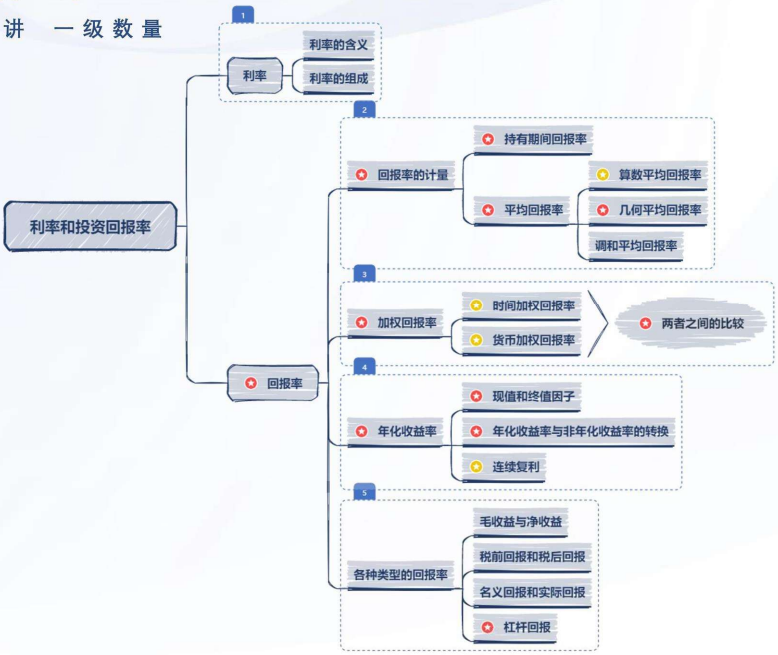
- Continuously compounded
  - $R^c_{annual} = e^{r_s} - 1$
  - $FV_N = PV_0 \times e^{r_s \times N}$
- $r^c_{t,t+1} = \ln(P_{t+1}/P_t) = \ln(1 + HPR_{t,t+1})$
- $r^c_{0,T} = r^c_{T-1,T} + r^c_{T-1,T-2} + \cdots + r^c_{0,1}$

5. 各类回报率

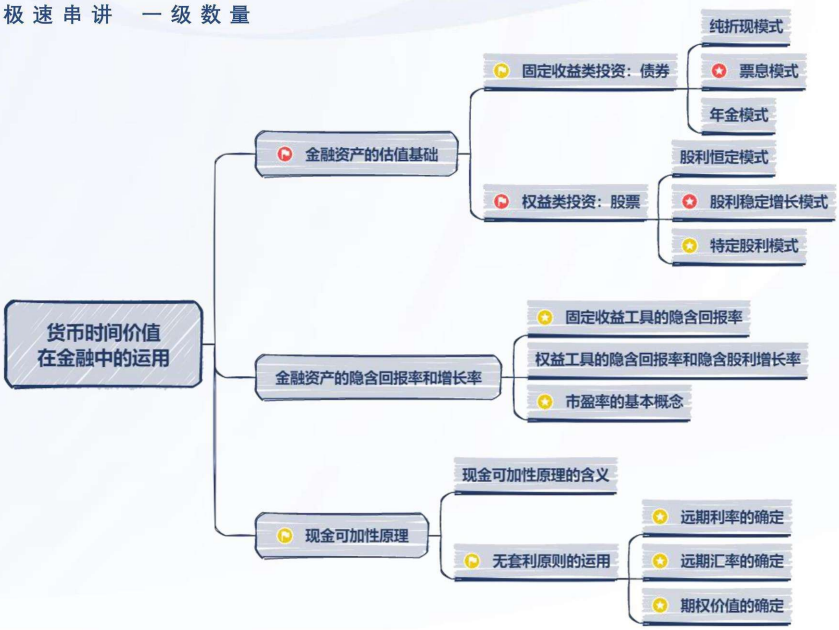
- **Gross return** is the return before deducting management expenses, custodial fees and taxes
  - A gross return is the return earned by an asset manager prior to deductions for the expenses that are not directly related to the generation of returns but rather related to the management and administration of an investment
- **Net return** is the return after deducting all fees and expenses
- The **after-tax real return** is what the investor receives as compensation for postponing consumption and assuming risk after paying taxes on investment returns
  - Many investors are concerned about the possible tax liability associated with their returns because taxes reduce the net return that they receive
  - 1 + real return = (1 + nominal risk free rate) × (1+ risk premium) / (1 + inflation premium)
- Via derivatives or borrowing, a **leveraged position** can be created, which may magnify the gains and losses
  - Leveraged return (R_L) = $r + v_B/v_C \times (r - r_B)$

王牌陈讲CFA
极速串讲 一级数量

一. 利率和投资回报率【1-5】



货币时间价值在金融中的应用



王牌陈 讲CFA
极速串讲 一级数量

高顿教育 GOLDEN EDUCATION

二

描述性统计

高顿教育 GOLDEN EDUCATION

6. 统计学的基本概念

- **Descriptive statistics** is a study of how data can be summarized effectively to describe the important aspects of data sets
  1. **Central tendency**: where data are centered
  2. **Dispersion**: how far data are dispersed from their center, usually used to address the risk
     - **Covariance** and **correlation**: how two variables move together
  3. **Skewness**: whether the distribution of data is symmetrically shaped
  4. **Kurtosis**: whether extreme outcomes are likely or whether fatty tails exist
- **Inferential statistics** involves making forecasts, estimates, or judgments about a larger group from the smaller group actually observed

- **Population** includes all members of a specified group
  - **Parameter** is the descriptive measure of a population characteristic
- **Sample** is a subset of a population
  - **Sample statistic** is a descriptive measure of a sample characteristic

7.1 均值

1. **Arithmetic mean** return focus on average single-period performance
   - ◆ population arithmetic mean $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$
   - ◆ sample arithmetic mean $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
   - ● Advantage: easy to work with mathematically; uses all information about the size and magnitude
   - ● Disadvantage: sensitive to extreme values, or outliers
   - ☐ When dealing with a sample that has extreme values
     - I. there may be a possibility of transforming the variable (e.g., a log transformation) or of selecting another variable that achieves the same purpose
     - II. Do nothing, use the data without any adjustment
     - III. Delete all the outliers, calculate **trimmed mean**
     - IV. Replace the outliers with another value, calculate **winsorized mean**
2. **Weighted mean**
3. **Geometric mean** return focus on an investment over a multi-period horizon
4. **Harmonic mean**

7.2 中位数和众数

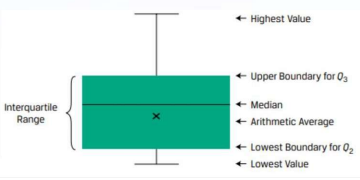
- ☐ **Median** is the value of the middle item of a set of ascending or descending order
  - ■ Odd number of n items, median occupies the (n+1)/2 position
  - ■ Even number of n items, median is equal to the mean of the items occupying the n/2 and (n+2)/2 positions
- ● Advantage: not affected by extreme values (outliers) as arithmetic mean
- ● Disadvantage: only one or two numbers considered, rest is to be ignored
- ☐ **Mode** is the most frequently occurring value of the distribution
  - ■ The distribution could have one mode (unimodal), more than one mode (bimodal, trimodal, etc.), or even no mode

7.3 分位数

- ☐ **Quantile** is a value at or below which a stated fraction of the data lies
  - ■ **Quartiles**: the distribution divided into quarters
    - The **interquartile range (IQR)** is the difference between the third quartile and the first quartile, or **IQR = $Q_3 - Q_1$**
  - ■ **Quintiles**: the distribution divided into the fifths
  - ■ **Deciles**: the distribution divided into the tenths
  - ■ **Percentiles**: the distribution divided into the hundredths
- ☐ Formula for location of data in ascending order
  - ◆ $L_y = (n+1) \times y/100$

    y is the y^{th} percentile

    n is the number of data
- ☐ **Box and whisker plot**



8.1 绝对离散程度

- **Dispersion** describes the variability around the central tendency
  - If mean return addresses reward, then dispersion addresses risk and uncertainty

1. **Range = maximum value – minimum value**
   - Only use two numbers and tell nothing about the distribution of the data set
2. **Mean absolute deviation** MAD = $\frac{\sum_{i=1}^{n}|X_i-\bar{X}|}{n}$
3. **Variance** is defined as the average of the squared deviations around the mean
   - population variance $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$
   - sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n-1}$
4. **Standard deviation** is the square root of the variance
   - population standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$
   - sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n-1}}$
   - The **target downside deviation** (**target semi-deviation**) is a measure of dispersion of the below the target
     - $s_{target} = \sqrt{\frac{\sum_{all\ x_i \le B}(x_i-B)^2}{n-1}}$

8.2 相对离散程度

- **Coefficient of variation (CV)** is the ratio of the standard deviation of a set of observations to their mean value, so it measures the amount of risk (standard deviation) per unit of reward (mean return)
  - CV has no units of measurement, so permits direct comparisons of dispersions across different data sets

9.1 散点图

- A **scatter plot** is a useful tool for displaying and understanding potential relationships between two variables
  - Tight (loose) clustering signals a potentially stronger (weaker) relationship between the two variables
  - Scatter plots are a powerful tool for finding patterns between two variables, for assessing data range, and for spotting extreme values

9.2 协方差

- **Covariance** is a measure of how two pair variables move together
  - Population covariance $Cov(X, Y) = 1/N \times \sum[(X_i - \mu_X)(Y_i - \mu_Y)]$
  - Sample covariance $s_{XY} = 1/(n-1) \times \sum[(x_i - \bar{X})(y_i - \bar{Y})]$
- It measures the linear relationship between two variables
  1. Positive covariance shows two variables *tend to* increase or decrease at same time
  2. Negative covariance shows one variable *tend to* increase when the other decreases
  3. Zero covariance means no linear relationship between two variables
- **Autocovariance** is equal to the variance
- The size of the covariance measure alone is difficult to interpret as it involves squared units of measure and so depends on the magnitude of the variables
  - Covariance values range from negative infinity to positive infinity

9. 3 相关系数

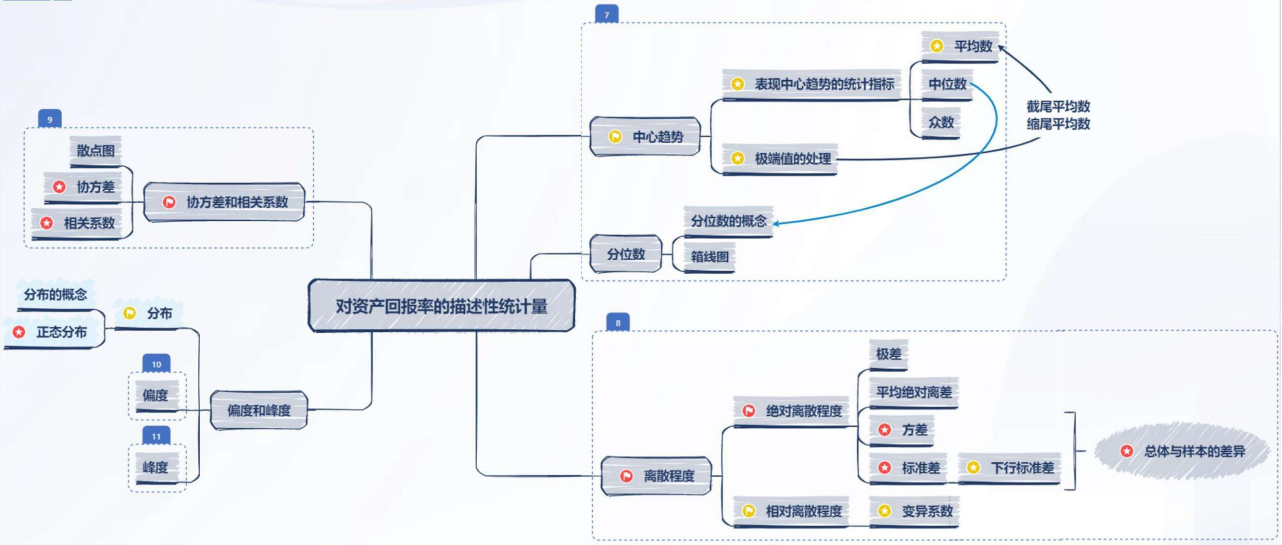
- The **correlation coefficient** is a standardized (normalized) measure, which expresses the strength of the linear relationship between two variables
  - Population correlation $\rho_{X,Y} = \sigma_{X,Y} / (\sigma_X \sigma_Y)$
  - Sample correlation $r_{X,Y} = s_{X,Y} / (s_X s_Y)$
- Correlation values range from +1 (perfect positive correlation) to –1 (perfect negative correlation)
- Limitations of correlation analysis
  - Two variables can have a strong non-linear relation and still have a very low correlation
  - The correlation may be quite sensitive to outliers
  - Correlation does not imply causation
  - The term **spurious correlation** has been used to refer to:
    - 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, but from their relation to a third variable

10. 偏度

- **Skewness** indicates the degree of symmetry of return distributions
  - Sample skewness $S_k \approx \frac{1}{n} \times \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$
  - $S_k = 0$ means symmetrical distribution: mode = median = mean
  - $S_k > 0$ means positively (right) skewed distribution: mode < median < mean
  - $S_k < 0$ means negatively (left) skewed distribution: mode > median > mean

11. 峰度

- **Kurtosis** measures the degree to which the tails of a distribution relative to the normal distribution
  - Sample kurtosis $k \approx \frac{1}{n} \times \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4}$
  - **Excess kurtosis** = kurtosis – 3
  - **Leptokurtic**: fatter tailed than normal distribution, kurtosis > 3, excess kurtosis > 0
  - **Mesokurtic**: identical to normal distribution, kurtosis = 3, excess kurtosis = 0
  - **Platykurtic**: thinner tailed than normal distribution, kurtosis < 3, excess kurtosis < 0



三 概率论基础

12. 随机变量的期望和方差

- **Random variable** is a quantity whose possible values are uncertain
  - **Outcomes** is the possible values of a random variable
- **Event** is a specified set of outcomes
- The **expected value of a random variable** is the probability-weighted average of the possible outcomes of the random variable
  - $E(X) = P(X=x_1)\cdot x_1 + P(X=x_2)\cdot x_2 + \cdots\cdots + P(X=x_N)\cdot x_N$
    $= P(x_1)\, x_1 + P(x_2)\, x_2 + \cdots\cdots + P(x_N)\, x_N = \Sigma\, P(x_i)\, x_i$
- The **variance of a random variable** is the expected value (probability-weighted average) of squared deviations from the random variable's expected value
  - $\sigma^2(X) = P(x_1)\cdot[x_1 - E(X)]^2 + \cdots\cdots + P(x_N)\cdot[x_N - E(X)]^2 = E[X - E(X)]^2$
- **Covariance** given a joint probability function
  - $Cov_{AB} = \Sigma_i\, P(R_A = a_i, R_B = b_i) \times [a_i - E(R_A)] \times [b_i - E(R_B)]$

13. 条件概率

- **Unconditional probability** (**marginal probability**) is the probability of an event A not conditioned on another event, denoted as **P(A)**
- **Conditional probability** is the probability of an event A conditioned on another event B, denoted as **P(A|B)**
- For independent events, **P(A|B) = P(A)**
  - If the occurrence of event A does not influence the occurrence of event B, then A and B are **independent**
- **Joint probability** is the probability of event A and B both happen, denoted **P(AB)**
  - $P(AB) = P(B)\times P(A|B) = P(A)\times P(B|A)$
- **Total probability rule** explains the unconditional probability of the event A in terms of probabilities conditional on the scenarios
  - $P(A) = P(A|S_1)\times P(S_1) + P(A|S_2)\times P(S_2) + \cdots\cdots + P(A|S_n)\times P(S_n)$
    where $S_1, S_2, \cdots\cdots S_n$ are mutually exclusive and exhaustive

14. 条件期望和条件方差

- **Conditional expected value** $E(X|S) = P(x_1|S)\,x_1 + P(x_2|S)\,x_2 + \cdots + P(x_N|S)\,x_N$
- **Conditional variance** $\sigma^2(X|S) = P(x_1|S)\,[x_1 - E(X|S)]^2 + \cdots + P(x_N|S)\,[x_N - E(X|S)]^2$
- **Total probability rule for expected value**

  $E(X) = E(X|S_1)P(S_1) + E(X|S_2)P(S_2) + \cdots\cdots + E(X|S_n)P(S_N)$

15. 贝叶斯法则

- Given a prior probabilities P(A) for an event of interest, if receive new information (B), **Bayes' formula** is the rule used for updating the probability (updated probability, posterior probability, P(A|B)) of the event
  - $P(A|B) = \frac{P(B|A)}{P(B)} \times P(A)$ $\quad P(A|B) = \frac{P(B|A)}{P(B)} \times P(A)$

16. 投资组合回报率的期望值和方差

- **Expected return on the portfolio** $E(R_p) = P(r_1)\,r_1 + P(r_2)\,r_2 + \cdots + P(r_m)\,r_m$

  $E(R_p) = E(w_1R_1 + \cdots + w_nR_n) = w_1E(R_1) + \cdots + w_nE(R_n)$
- **Portfolio variance of return** $\sigma^2(R_p) = E[(R_p - E(R_p))^2] = \Sigma_{i=1}\Sigma_{j=1}w_iw_jCov(R_i, R_j)$

1. If the portfolio consists of two assets (1 and 2)
   - $E(R_p) = w_1E(R_1) + w_2E(R_2)$ $\quad \sigma_p^2 = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_1\sigma_2\rho_{1,2}$
2. If the portfolio consists of three assets (1 and 2 and 3)
   - $E(R_p) = w_1E(R_1) + w_2E(R_2) + w_3E(R_3)$
   - $\sigma_p^2 = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + w_3^2\sigma_3^2 + 2w_1w_2\sigma_1\sigma_2\rho_{1,2} + 2w_1w_3\sigma_1\sigma_3\rho_{1,3} + 2w_2w_3\sigma_2\sigma_3\rho_{2,3}$

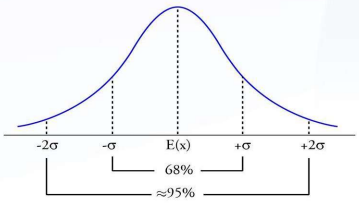
17. 1 正态分布

- **Normal distribution** (**Gaussian distribution**), which is a symmetrical, bell-shaped distribution, plays a central role in the mean–variance model of portfolio selection; it is also used extensively in financial risk management
  - It is completely described by two parameters: mean and variance $X \sim N(\mu, \sigma^2)$
  - Symmetric distribution, and its mean, median, and mode are equal
  - Probabilities decrease further from the mean, but the tails go on forever
- **68%** confidence interval = $\mu \pm$ **1**σ

  90% confidence interval = $\mu \pm$ **1.65**σ

  95% confidence interval = $\mu \pm$ **1.96**σ

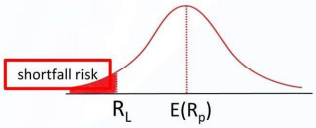
  99% confidence interval = $\mu \pm$ **2.58**σ



17. 2 标准正态分布

- **Standard normal distribution** (**Z-distribution**) is the normal distribution with mean $\mu = 0$, and standard deviation $\sigma = 1$
- **Standardization**: if $X \sim N(\mu, \sigma^2)$, then $\mathbf{Z = (X - \mu) / \sigma \sim N(0, 1)}$
- Z-table: $\Phi(k) = P(X \leq k)$, $\Phi(-k) = 1 - \Phi(k)$

17. 3 安全第一比率和亏空风险



- **Shortfall risk** is the risk that portfolio value or return will fall below the minimum acceptable level (R_L, shortfall level, threshold level) over some time horizon
  - **Safety-first ratio (SFRatio) = $[E(R_p) - R_L] / \sigma_p$**
- The portfolio for which $E(R_p) - R_L$ is largest relative to standard deviation minimizes $P(R_p < R_L)$
  - $P(R_p \leq R_L) = \Phi(-SFRatio)$

17. 4 对数正态分布

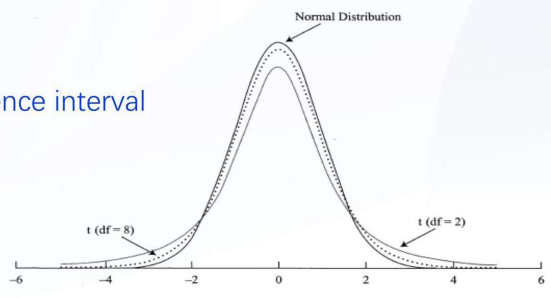
- If **X** is **normally distributed**, then $\mathbf{e^x}$ is **lognormal distributed**
  - If **Y** follows a **lognormal distribution**, then **lnY** is **normally distributed**
- Lognormal random variable is bounded from below by zero, and it is positively skewed
- The stock price may be well described by the lognormal distribution, and a stock's continuously compounded return is normally distributed

17. 5 连续复利模式下的回报率

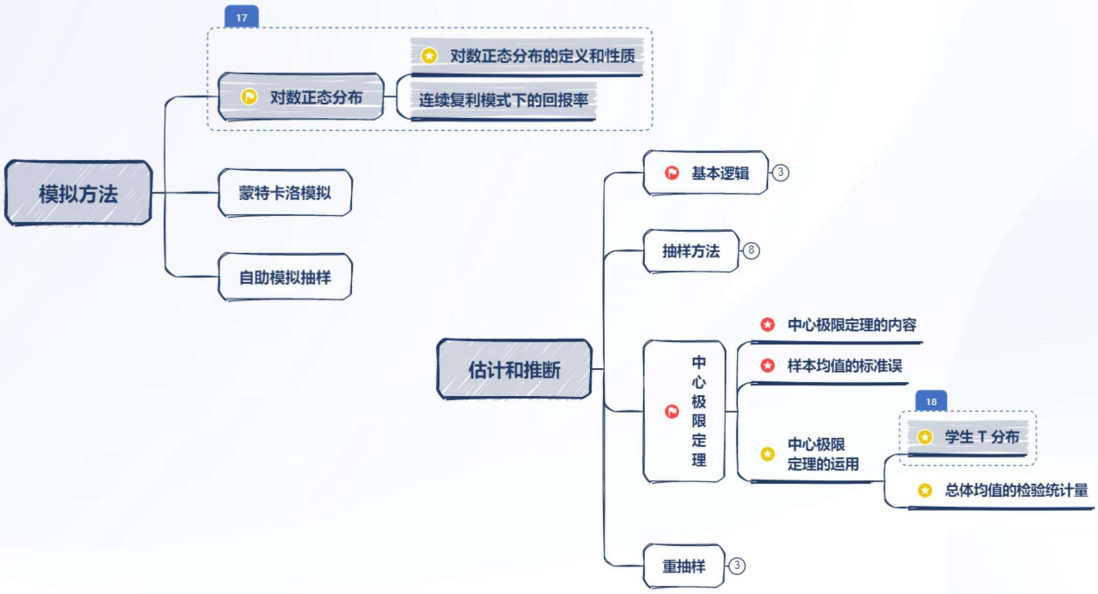
- If continuously compounded returns are independently and identically distributed (i.i.d.)
  - $\mathbf{E(r_{0,T}) = E(r_{T-1,T}) + E(r_{T-1,T-2}) + \cdots + E(r_{0,1}) = \mu T}$
  - $\mathbf{\sigma^2(r_{0,T}) = \sigma^2 T \quad \Rightarrow \quad \sigma(r_{0,T}) = \sigma\sqrt{T}}$

18. 学生 T 分布

- **Student's T-distribution** is defined by single parameter: degrees of freedom (df)
  - df = n – 1, where n is the sample size
  - Symmetrical (bell shaped), skewness = 0
  - Fatter tails than a normal distribution
  - As df increase, T-distribution is approaching to standard normal distribution
- Given a degree of confidence, T-distribution has a wider confidence interval than Z-distribution



王牌陈 讲CFA
极速串讲 一级数量
三. 概率论基础【12－18】
高顿教育 GOLDEN EDUCATION



王牌陈 讲CFA
极速串讲 一级数量
高顿教育 GOLDEN EDUCATION

四

抽样和推断

19. 抽样的基本逻辑

- **Sampling error** is the difference between the sample statistic (which is a random variable) and the population parameter (which is a constant)
- **Sampling distribution** is the distribution of all distinct possible values that the statistic can assume when computed from samples of the same size randomly drawn from the same population

20. 抽样方法

1. Using **probability sampling**, every member of the population has an equal chance being selected
   - Probability sampling can create a representative sample of the population, so it is more accuracy and reliability
     ① **Simple random sampling**: each element of the population has an equal probability of being selected to the subset, and it is useful for homogeneous data

20. 抽样方法

1. **Probability sampling**
   ② **Stratified random sampling**: **1)** separate the population into subpopulations based on one or more classification criteria, **2)** use simple random sampling to draw from each stratum in sizes proportional to the relative size of each stratum in the population and then pooled to form a stratified random sample
   ③ **Systematic sampling**: **1)** randomly pick a member as the starting number, **2)** select every k^{th} member until reaching the sample size
   ④ **Cluster sampling**: **1)** divide the population into clusters based on some criteria, **2)** use simple random sampling to choose certain clusters
   - One-stage cluster sampling: all the members in each sampled cluster are included
   - Two-stage cluster sampling: a subsample is selected from each sampled cluster through random sampling

20. 抽样方法

2. **Non-probability sampling** depends on factors other than probability, and it may generate a non-representative sample

   ① **Convenience sampling**
   - An element is selected based on convenience of accessibility
   - Advantage: time-efficient and cost-effective
   - Disadvantage: limited level of accuracy

   ② **Judgmental sampling**
   - Elements are selected based on a sampler's specialty and professional judgment
   - Results may be skewed due to the bias of the sampler

21. 蒙特卡洛模拟

- **Monte Carlo simulation** uses randomly generated values for risk factors, based on their assumed distributions, to produce a distribution of possible outcomes
  - Monte Carlo simulation is widely used to estimate risk and return investment applications, and it is often combined with scenario analysis or sensitivity analysis
  - 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
  - Monte Carlo simulation is a complement to analytical methods, and it provides only statistical estimates, not exact results
    - Analytical methods provide more insight into cause-and-effect relationships

22. 重抽样

- **Resampling** is a computational tool to repeatedly draw samples from the original sample for the statistical inference of population parameters

1. **Bootstrap** is one of the most popular resampling methods, which repeatedly draws samples from the original sample, and each resample is of the same size with the original sample
   - The identical element is put back into the group so that it can be drawn more than once
   - Bootstrap resampling is model-free (non-parametric resampling), and it does not rely on an analytical formula
     - It is a complement to analytical methods, so it provides only statistical estimates
2. **Jackknife** is another resampling method, which will leave out one observation from the original sample at one time without replacing it, thus usually requires repetitions equal to the sample size
   - Jackknife resampling is used to reduce the bias
   - Jackknife resamples produce similar results, whereas bootstrap resamples usually have different results

23. 1 中心极限定理

- **Central Limit Theorem**: given a population described by **any** probability distribution **having** mean μ and finite variance σ^2, the sampling distribution of the **sample mean** $\bar{X}$ computed from random samples of size n from this population will be approximately normal with mean μ (population mean) and variance σ^2/n (population variance divided by n) when the sample size n is large
  - The central limit theorem helps us understand the sampling distribution of the sample mean, and it is used to estimate how closely the sample mean can be expected to match the underlying population mean

23. 2 样本均值的标准误

- The standard deviation of a sample statistic is known as the **standard error of the statistic**
  1. When the population standard deviation (σ) is known: $\sigma_{\bar{X}} = \sigma / \sqrt{n}$
     - $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$
  2. When the population standard deviation (σ) is unknown: $s_{\bar{X}} = s / \sqrt{n}$
     - $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim T_{(df = n-1)} \overset{\text{when n is large}}{\sim} N(0, 1)$
- Although the standard error is the standard deviation of the sampling distribution of the parameter, "standard deviation" and "standard error" are two distinct concepts
  - Standard deviation measures the dispersion of the data from the mean
  - Standard error measures how much inaccuracy of a population parameter estimate comes from sampling

24. 1 假设检验的基本概念

- **Hypothesis testing** is an act in statistics whereby an analyst tests an assumption regarding a population parameter
- Decisions: either **reject** the hypothesis, or **not** reject the hypothesis

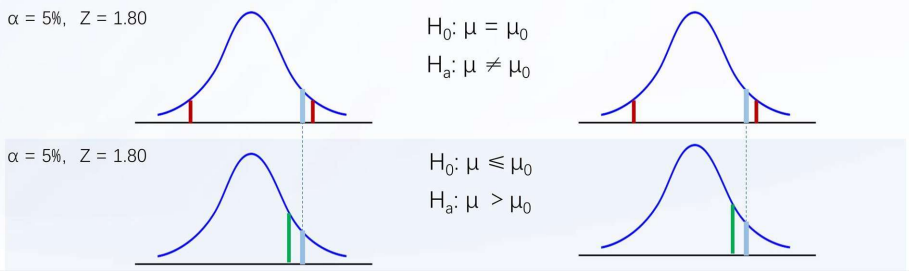
24. 2 原假设和备择假设

- **Null hypothesis (H_0)** are hypothesis to be tested
- **Alternative hypothesis (H_a)** are the opposite side of null hypothesis
  - The null and alternative hypotheses must be mutually exclusive and collectively exhaustive

1. **Two-tailed test** $H_0: \theta = \theta_0$ vs. $H_a: \theta \neq \theta_0$
2. **One-tailed test** $H_0: \theta \leqslant \theta_0$ vs. $H_a: \theta > \theta_0$ ($H_0: \theta \geqslant \theta_0$ vs. $H_a: \theta < \theta_0$)

24. 3 假设检验的决策依据

- **Sample statistic** & **Test statistic**
- **Significance level (α)** & **Confidence level (1 – α)**
- If the calculated value of the test statistic is within rejection region (critical region), then we reject the null hypothesis, and we say the result is statistically significant
  1. If test statistic's value is outside the range of critical value
     | test statistic's value | > | critical value | ⇨ reject the null hypothesis
  2. If the **P-value < level of significance (α)** ⇨ reject the null hypothesis
     - **P-value** is the smallest level of significance at which the null hypothesis can be rejected



25. 一类错误和二类错误

- **Type I error** is rejecting null hypothesis when it is true
  - **P(**type I error**)** = **P(**reject null hypothesis | the null is true**)** = level of significance (α)
- **Confidence level** (1 – α) = 1 – P(Type I Error)
  = **P(**not reject null hypothesis | the null is true**)**
- **Type II error** is failing to reject the null hypothesis when it is false
  - **P(**type II error**)** = **P(**not reject the null | the null is false**)** = β
- **Power of test** = 1 – P(Type II Error) = 1 – β
  = **P(**reject the null | the null is false**)**
- When sample size increases, the probabilities of both Type I error and Type II error decrease

25-1

26. 1 对单一均值的检验

- Test statistic for hypothesis tests of population mean with known variance
  - $\mathbf{z} = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$
- Test statistic for hypothesis tests of population mean with unknown variance
  - $\mathbf{T} = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \sim T_{n-1} \overset{\text{when n is large}}{\sim} N(0, 1)$

26. 2 对多个均值的检验

- Hypothesis test concerning difference of means
  - $H_0: \mu_1 - \mu_2 = d_0$
  - Independent populations with variance unknown but assumed equal
  - $\mathbf{T} = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim T_{n_1+n_2-2}$
    $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$
- Hypothesis test concerning the mean of the differences (a paired comparisons test)
  - $H_0: \mu_d = d_0$
  - Dependent populations
  - $\mathbf{T} = \frac{\bar{d} - d_0}{s_{\bar{d}}} = \frac{\frac{1}{n}\sum d_i - d_0}{s_d/\sqrt{n}} \sim T_{n-1}$

26-1

27. 1 对单一方差的检验

- Hypothesis test concerning a single variance
  - $H_0: \sigma^2 = \sigma_0^2$
  - $\chi^2 = \frac{(n-1)\times s^2}{\sigma_0^2} \sim \chi^2_{n-1}$

27. 2 对多个方差的检验

- Hypothesis test concerning equality of two variances
  - $H_0: \sigma_1^2 = \sigma_2^2$
  - Independent populations
  - $\mathbf{F} = s_1^2/s_2^2 \sim F(n_1-1, n_2-1)$

27-1

28. 参数检验和非参数检验

- A **parametric test** is concerned with parameters
  - The validity depends on a definite set of assumptions — in particular, assumptions about the distribution of the population producing the sample
- A **nonparametric test** is a test that is not concerned with a parameter or a test that makes minimal assumptions about the population from which the sample comes
  - The parametric test (where available) is generally preferred over the nonparametric test because the parametric test may have more power — that is, a greater ability to reject a false null hypothesis
  - Nonparametric test is applied when:
    - Data do not meet distributional assumptions
    - There are outliers
    - Data are given in ranks or use an ordinal scale
    - The hypothesis we are addressing does not concern a parameter

29. 1 对相关系数的参数检验

- Parametric correlation refers to **Pearson correlation**, the bivariate correlation, or simply the correlation
  - $\rho_{XY} = Cov(X,Y) / (\sigma_X \sigma_Y)$
  - $H_0: \rho = 0$ (or $H_0: \rho \geq 0$ or $H_0: \rho \leq 0$)
  - $T = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}} \sim T_{n-2}$

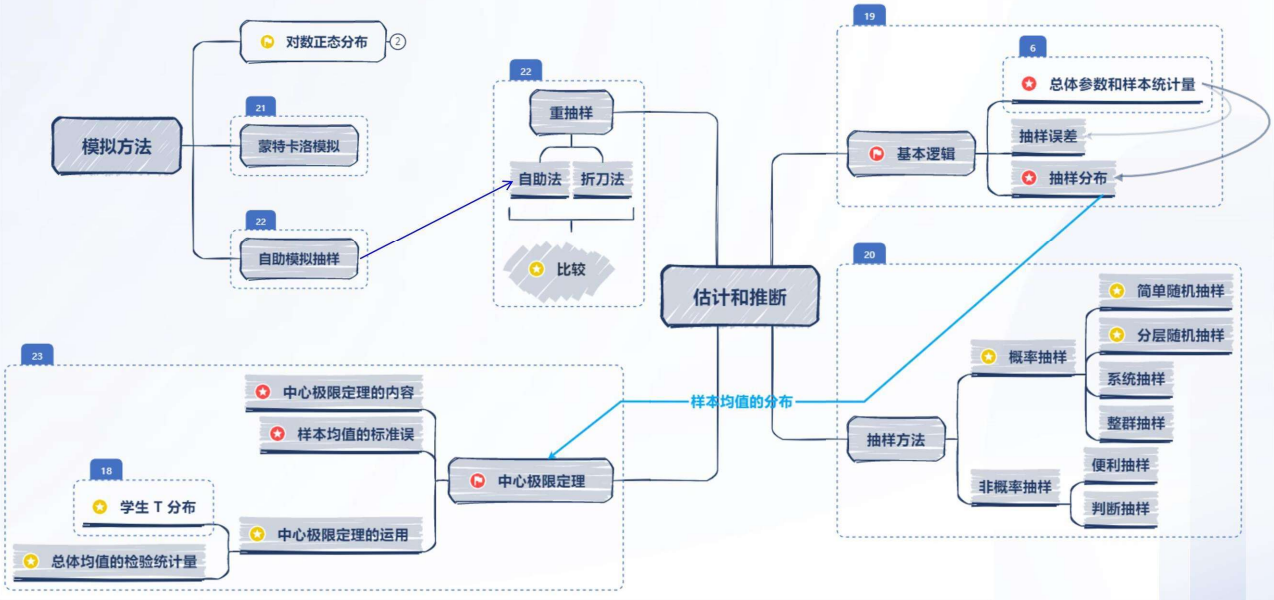
29. 2 对相关系数的非参数检验

- Nonparametric test of a correlation is based on the **Spearman rank correlation coefficient** between variables X and Y
  - $r_s = 1 - \frac{6 \times \sum_{i=1}^{n} d_i^2}{n \times (n^2-1)}$
  - $d_i = rank_{X_i} - rank_{Y_i}$
  - The test of hypothesis for the Spearman rank correlation depends on whether the sample is small or large (n>30)
    - For small samples, the researcher requires a specialized table of critical values
    - For large samples, we can conduct a t-test with n − 2 degrees of freedom

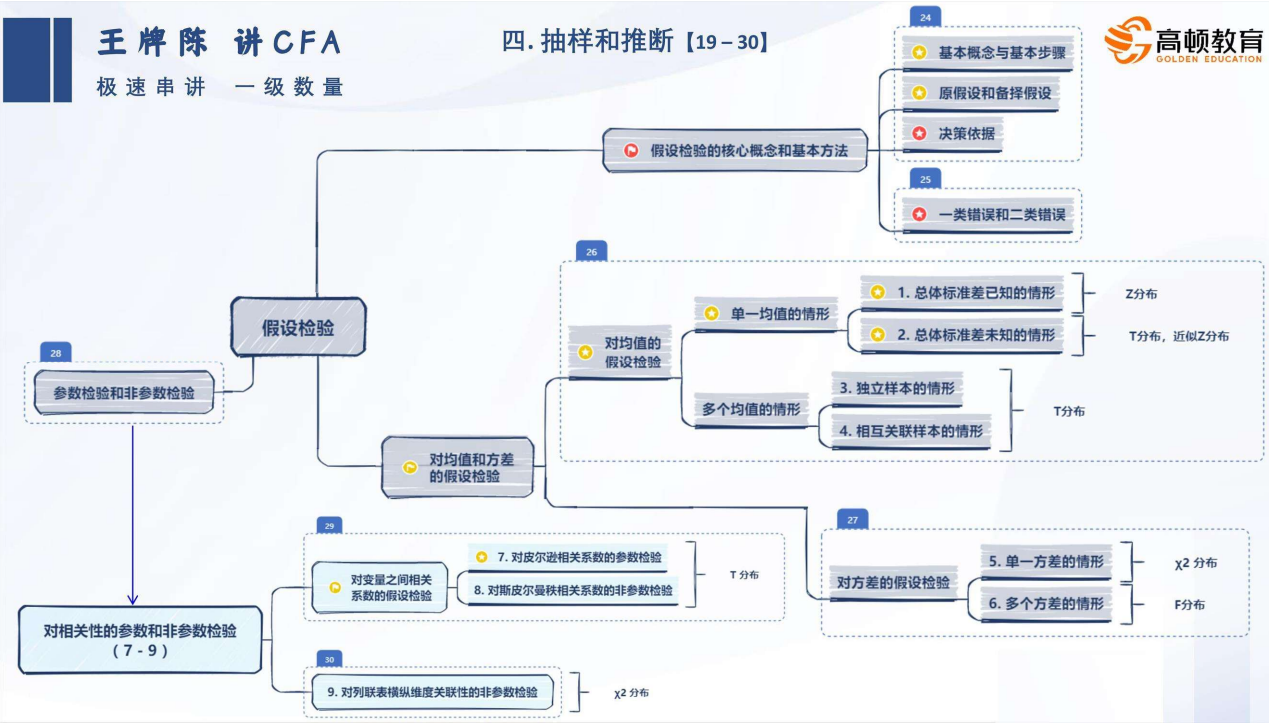
30. 对列联表不同维度关联性的非参数检验

- Tests of independence based on **contingency table** data is to test whether the classifications of categorical or discrete data are independent
  - It is a nonparametric one-tailed chi-square test
  - $\chi^2 = \sum_{i=1}^{m} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)\times(c-1)}$
    - m is the number of cells in the contingency table (m = i × j)
    - O_{ij} is the observed frequency in the $cell_{(i,j)}$
    - E_{ij} is the expected frequency in the $cell_{(i,j)}$

四. 抽样和推断【19－30】



四. 抽样和推断【19－30】



五 简单线性回归

31. 简单线性回归的基本概念

- **Linear regression** shows the linear relationship between dependent and independent variables
  - The variable whose variation is being explained as **dependent variable** (**explained variable**)
  - The variable(s) whose variation is (are) being used to explain the variation of the dependent variable as **independent variable(s)** (**explanatory variable(s)**)
- If we have only one independent variable, we refer this as **simple linear regression (SLR)**
- Population regression function $Y_i = b_0 + b_1X_i + \varepsilon_i \quad i = 1, \cdots, n$
  - **Population parameters b_0** and **b_1** are the **intercept** and the **slope coefficient**
  - ε_i is the **error term**, **$E(\varepsilon) = 0$**
- Sample regression function $Y_i = \hat{b}_0 + \hat{b}_1X_i + e_i \quad \hat{Y}_i = \hat{b}_0 + \hat{b}_1X_i \quad i = 1, \cdots, n$
  - $\hat{b}_0$ and $\hat{b}_1$ are the **estimates of the population parameters**
    - $e_i = Y_i - \hat{Y}_i$

32. 最小二乘法

- In simple linear regression, the estimated intercept $\hat{b}_0$, and slope $\hat{b}_1$, are such that sum of the squared vertical distances from the observations to the fitted line is minimized
  - Fitting the line requires minimizing the **sum of squares error (SSE)**, also known as the **residual sum of squares**
    - $SSE = \Sigma (Y_i - \hat{Y}_i)^2 = \Sigma [Y_i - (\hat{b}_0 + \hat{b}_1 X_i)]^2 = \Sigma e_i^2$
- The slope ($\hat{b}_1$) is the ratio of the covariance between Y and X to the variance of X
  - $\hat{b}_1 = \frac{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{n-1}}{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- Once estimate the slope ($\hat{b}_1$), we can then estimate the intercept ($\hat{b}_0$) using the mean of Y and the mean of X
  - $\hat{b}_0 = \bar{Y} - \hat{b}_1 \times \bar{X}$

33. 简单线性回归的前提假设

- **Assumption 1**: the relationship between the dependent variable Y and the independent variable X is linear
  - When we look at the residuals of a model, what we would like to see is that the residuals are random, and the independent variable X is not necessarily random
- **Assumption 2**: the variance of the regression residuals is the same for all observations, which is known as the **homoskedasticity**
  - **Heteroskedasticity**: the variance of residuals differs across observations
- **Assumption 3**: the observations, pairs of Y_i and X_i, are independent of one another, which implies the regression residuals are independent across observations
- **Assumption 4:** the regression residuals are normally distributed
  - This does not mean that the dependent and independent variables must be normally distributed

34. 1 方差分析表

Analysis of Variance Table for Simple Linear Regression				
Source	Sum of Squares	Degrees of Freedom	Mean Square	
Regression	SSR = $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	1	MSR = SSR / 1	
Error	SSE = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	n – 2	MSE = SSE / (n–2)	$S_e = \sqrt{MSE}$
Total	SST = $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	n – 1	Var(Y) = SST / (n–1)	
	R^2 = SSR / SST		F = MSR / MSE	

34. 2 拟合优度评价

- **Coefficient of determination (R^2)** represents the percentage of the variation of the dependent variable that is explained by the independent variable
  - ◆ **R^2 = SSR / SST = 1 – SSE / SST**
  - R^2 ranges from 0% to 100%
  - In a simple linear regression, the square of the pairwise correlation is equal to the coefficient of determination, **$R^2 = r^2$**
- **F-statistics F = MSR / MSE = (SSR/1) / [SSE/(n–2)]**
  - $H_0 : b_1 = 0$ $H_a : b_1 \neq 0$
  - It is distributed with 1 and n – 2 degrees of freedom in simple linear regression
- **Standard error of the estimate $S_e = MSE^{0.5}$**
  - The smaller the S_e, the better the fit of the model

35. 1 对参数的假设检验

- Hypothesis tests of the slope coefficient ($H_0 : b_1 = M$)
  - $t = \frac{\hat{b}_1 - M}{s_{\hat{b}_1}} \sim T_{n-2}$
  - **Standard error of the slope coefficient** $s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
    - The greater the variability of the independent variable, the lower the standard error of the slope and hence the greater the calculated t-statistic, and the null hypothesis is more likely to be rejected
- Hypothesis tests of the correlation ($H_0 : \rho = 0$)
  - $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim T_{n-2}$
  - Feature of simple linear regression: $t^2 = F$
- Hypothesis tests of the intercept ($H_0 : b_0 = N$)
  - $t = \frac{\hat{b}_0 - N}{s_{\hat{b}_0}} \sim T_{n-2}$

35. 2 对预测结果的估计区间

- Prediction interval of dependent variable : $\hat{Y}_f \pm t_{\text{critical for } \alpha/2} \times S_f$
  - **Standard error of the forecast** $s_f = s_e \times \sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- The smaller standard error of the forecast (S_f) will be achieved, if:
  - The better the fit of the regression model, the smaller the standard error of the estimate (S_e)
  - The larger the sample size (n) in the regression estimation
  - The larger the variation of the independent variable ($\sum_{i=1}^{n}(X_i - \bar{X})^2$)
  - The closer the forecasted independent variable (X_f) is to the mean of the independent variable ($\bar{X}$) used in the regression estimation

36. 1 指示变量

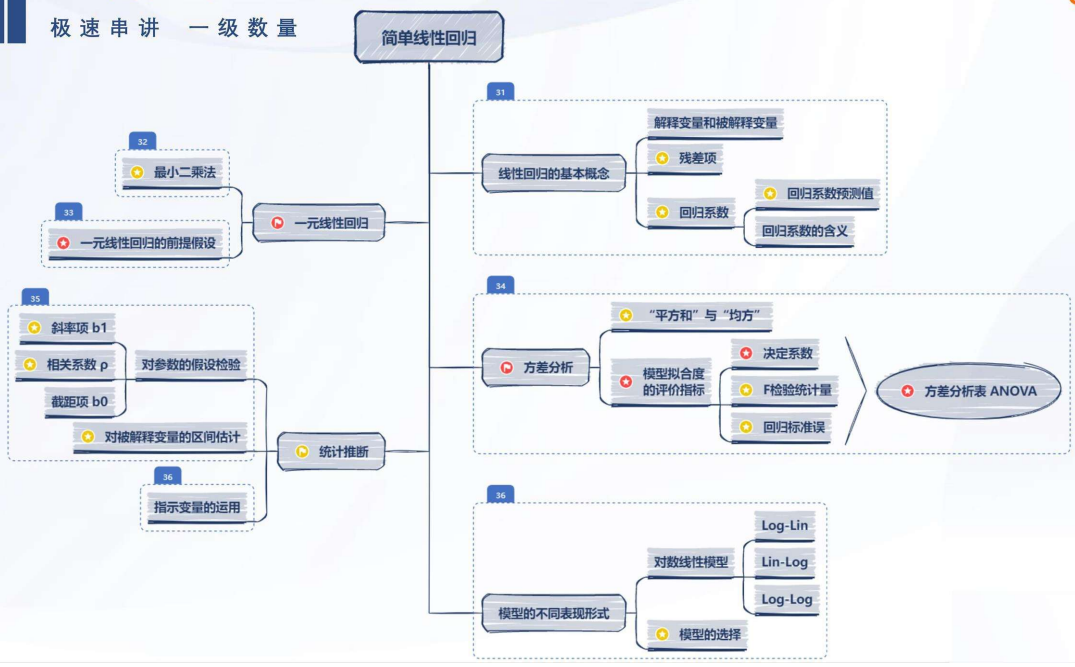
- **Indicator variable** (**dummy variable**) is an independent variable whose value takes on only 0 or 1
  - In a simple linear regression, the interpretation of the intercept (b_0) is the predicted value of the dependent variable if the indicator variable is 0
  - The interpretation of the slope (b_1) is the difference of the predicted value of the dependent variable if the indicator variable is 0 and 1, respectively

36. 2 对数线性回归模型

- **Log-lin model** $\ln Y_i = b_0 + b_1 X_i + \varepsilon_i$
  - The slope coefficient b_1 is the relative change in the dependent variable for an absolute change in the independent variable
- **Lin-log model** $Y_i = b_0 + b_1 \ln X_i + \varepsilon_i$
  - The slope coefficient b_1 provides the absolute change in the dependent variable for a relative change in the independent variable
- **Log-log model** $\ln Y_i = b_0 + b_1 \ln X_i + \varepsilon_i$
  - The slope coefficient b_1 is the relative change in the dependent variable for a relative change in the independent variable

王牌陈讲CFA 极速串讲 一级数量

五. 简单线性回归【31－36】



六 大数据分析

37. 金融科技的基本概念

- **Fintech** refers to technological innovation in the design and delivery of financial services and products
  - Areas of development that are more directly relevant to quantitative analysis in the investment: 1) analysis of large datasets, 2) analytical tools

38. 1 大数据的特征

1. **Volume**: MB ⇨ GB ⇨ TB ⇨ PB
2. **Velocity**: batch ⇨ periodic ⇨ real-time or near-real-time
3. **Variety**: structured data (e.g. SQL tables), semi-structured data (e.g. HTML), and unstructured data (e.g. videos)
4. **Veracity**: credibility and reliability of different data sources

38. 2 大数据的来源

- **Traditional data**: stock exchanges, financial statements, economic indicators
- **Non-traditional data (alternative data)**
  - Individuals: social media, news, reviews, web searches
  - Business processes: transaction data, corporate data
  - Sensors: satellites, geolocation, Internet of Things

38. 3 大数据分析面临的挑战与困境

- Quality, volume, and appropriateness of the data
- Unstructured data are more often qualitative
- Artificial intelligence and machine learning techniques have emerged

39.1 人工智能和机器学习

- **Artificial intelligence (AI)** enables the development of computer systems that exhibit cognitive and decision-making ability comparable or superior to that of human beings
  - **Neural networks**: based on brain learns and processes information to detect abnormal charges or claims in credit card fraud detection systems
- **Machine learning (ML)** is computer-based techniques that try to extract information from huge amounts of data without making any assumptions on the probability distribution of data
  - ML involves splitting the dataset into training dataset, validation dataset and test dataset

39.2 机器学习的分类

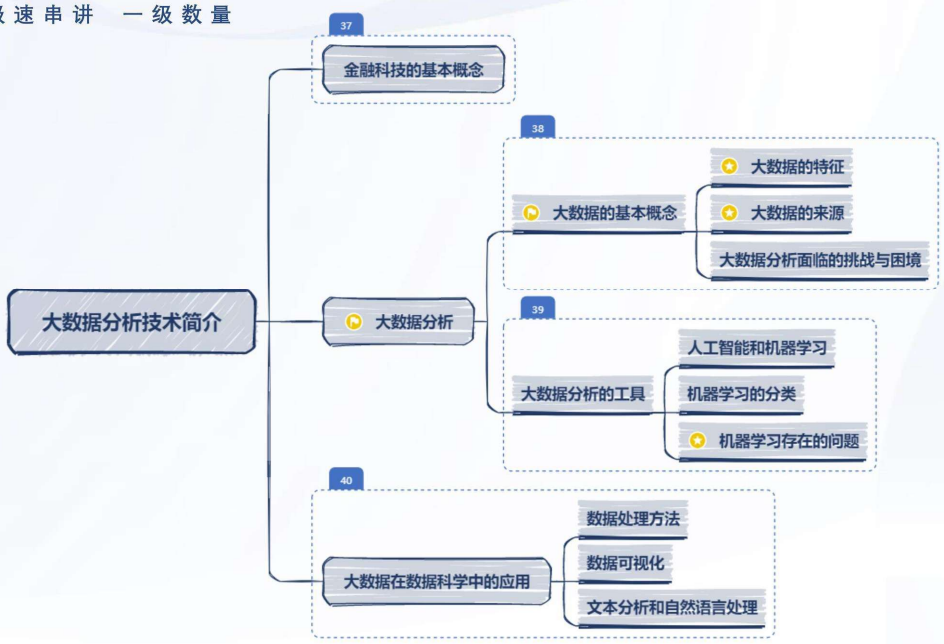
- **Supervised learning**: learn model relationships based on labeled training data
- **Unsupervised learning**: computers are not given labeled data, entailing the algorithm seeking to describe the data and their structure
- **Deep learning** (or deep learning nets): use neural networks, often with many hidden layers, to perform multistage, non-linear data processing to identify patterns

39.3 机器学习存在的问题

- ML still requires human judgment in understanding the underlying data and selecting the appropriate techniques for data analysis
- ML models also require sufficiently large amounts of data and might not perform well when not enough available data are available to train and validate the model
- **Overfitting** occurs when the model learns the dataset too precisely and treats noise as true
  Underfitting occurs when the model is too simplistic and treats true parameters as noise

40. 大数据在数据科学中的应用

- **Data visualization** refers to how the data will be formatted, displayed, and summarized in graphical form
  - Traditional structured data: tables, charts, and trends
  - Non-traditional unstructured data: three-dimensional (3D) graphics, tag cloud
- **Text analytics** is based on analyzing word frequency in a document to help identify future performance, such as consumer sentiment or economic trend
- **Natural language processing (NLP)** is a field of research at the intersection of computer science, artificial intelligence, and linguistics that focuses on developing computer programs to analyze and interpret human language



王牌陈 讲CFA

不要假装很努力
那些感动自己
结果不会陪你演戏