1. Quantitative

1.1. Interest Rate

1.1.1. 重要知识点

1.1.1.1. Interest rate

➢ Interest rate (r) measures the amount of money (interest) per unit of investment (principal).

➢ A rate of return that reflects the relationship between differently dated cash flows.

1.1.1.2. Decompose required rate of return

➢ Interest rate = real risk free rate + expected inflation rate + risk premium

➢ Nominal risk free rate = real risk-free rate + expected inflation rate

1.1.1.3. Simple and compounding interest rate

➢ **Simple interest rate (360 days):** there is no opportunity to re-invest the interest payments during the life of an investment and thereby earn extra income.

➢ **Compound interest rate (365 days):** Interest amounts will be received periodically and can be re-invested (usually at the same rate).

1.1.1.4. Types of interest

➢ **Required rate of return is**

■ affected by the **supply and demand** of funds in the market;

■ the minimum rate of return an investor must receive to accept the investment.

■ usually for particular investment.

➢ **Discount rate is**

■ the interest rate we use to **discount** payments to be made in the future.

■ usually used interchangeably with the interest rate.

➢ **Opportunity cost is**

■ also understood as a form of interest rate. It is the value that investors **forgo** by choosing a particular course of action.

1.1.2. 基础题

Q-1. An investor is considering purchasing short-term corporate bonds with similar maturities to US Treasury bills. If the real risk-free rate of US Treasury bill is 0.5%, the inflation premium is 1.5%, the default risk premium is 1% and the term risk premium is 2%, what is the approximate opportunity cost of not making this investment?

A. 2%

B. 3%

C. 5%

1.2. Calculation of HPY, EAR

1.2.1. 重要知识点

1.2.1.1. HPY 和 EAY/EAR 的计算及转化

➤ $HPR = \dfrac{FV - PV}{PV}$

➤ Effective annual rate of return (EAY/EAR)

■ $EAY = (1 + HPY)^{365/t} - 1$

■ $EAY = (1 + \dfrac{r}{m})^m - 1$

■ 性质：$r \uparrow$，$m \uparrow$，$EAR \uparrow$；When m $\to \infty$, $EAR_{max} = e^r - 1$

➤ 单利时用 $\dfrac{360}{t}$，复利时用 $\dfrac{365}{t}$

1.2.2. 基础题

Q-2. An investor is evaluating two fixed-term investments with the following features:

	Term Deposit 1	Term Deposit 2
Compounding frequency	continuous	monthly
Stated annual rate		6%

What annual interest rate would make the investor indifferent between investing in Term Deposit 1 and Term Deposit 2?

A. 5.83%

B. 5.99%

C. 6.18%

1.3. Time-Weighted Return and Money-Weighted Return

1.3.1. 重要知识点

1.3.1.1. 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.

1.3.1.2. 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.

1.3.1.3. 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.

1.3.2. 基础题

Q-3. An investor purchases 100 shares of a stock at time 0 for $400 per share. At times 1 and 2, the investor sells 50 shares each time, receiving $20,100 from both transactions. What is the Money-Weighted Rate of Return (MWRR) for this investment?

A. 0.50%

B. 0.45%

C. 0.33%

Q-4. Consider two series of returns over four periods, with each period representing one year:

	Year 1	Year 2	Year 3	Year 4
Series 1 (%)	-6	6	-6	6
Series 2 (%)	-11	11	-11	11

Which series of returns has a higher Time-Weighted Rate of Retun (TWRR)?

A. Set 1 has a higher TWRR.

B. Set 2 has a higher TWRR.

C. Both series have the same TWRR.

1.4. Time Value of Money

1.4.1. 重要知识点

1.4.1.1. Annuities 年金：FV, PV, required interest, payment

➢ N = number of periods

➢ I/Y = interest rate per period

➢ PMT = amount of each periodic payment

➢ FV = future value

> - PV = present value
> - 考察方法：计算——N, I/Y, PMT, FV, PV 中任意给定四个，求另外一个

1.4.1.2. Ordinary annuity 后付年金

> - The first cash flow occurs at the end of the first period (t=1)

1.4.1.3. Annuity due 先付年金 (BGN mode)

> - The first cash flow occurs immediately (at t=0)

1.4.1.4. Perpetuity

> - A perpetuity is a set of level never-ending sequential cash flows, with the first cash flow occurring one period from now.
>
> - 计算：$PV = \dfrac{A}{r}$

1.4.2. 基础题

Q-5. As winning a lottery, Mikey has three options to get bonus.

Option 1: An ordinary annuity with 20 annual payments of $2,000.

Option 2: An annuity due with 20 annual payments of $2,000.

Option 3: A perpetuity with annual payments of $2,000.

Assuming the annual discount rate is 5 percent, which option is the last one for Mikey to choose?

A. Option 1.

B. Option 2.

C. Option 3.

Q-6. For planning purposes, an individual wants to be able to spend €80,000 per year, at the end of each year, for an anticipated 20 years in retirement. In order to fund this retirement account, he will make annual deposits of €11,606.56 at the end of each of his working years. What is the minimum number of such deposits he will need to make to fund his desired retirement? Use 6% interest compounded annually for all calculations.

A. 29 payments

B. 30 payments

C. 31 payments

1.5. Measures of Central Tendency

1.5.1. 重要知识点

1.5.1.1. Median

> - Odd numbers: the (n+1)/2 position

> ➢ Even numbers: the n/2 position

1.5.1.2. Mode

> ➢ Unimodal; bimodal & Trimodal.

1.5.1.3. Measures of mean

> ➢ The arithmetic mean: $\overline{X} = \dfrac{\sum\limits_{i=1}^{N} X_i}{n}$

> ➢ The weighted mean: $\overline{X_W} = \sum\limits_{i=1}^{n} w_i X_i = (w_1 X_1 + w_2 X_2 + \cdots + w_n X_n)$

> ➢ The geometric mean: $G = \sqrt[N]{X_1 X_2 X_3 ... X_N} = (\prod\limits_{i=1}^{N} X_i)^{1/N}$

> ➢ The harmonic mean: $\overline{X_H} = \dfrac{n}{\sum\limits_{i=1}^{n} (1 / X_i)}$

> ➢ Harmonic mean ≤ geometric mean ≤ arithmetic mean

1.1.1.1. Performance measurement with means

> ➢ The geometric mean of past annual return is the appropriate measure of past performance.

> ➢ The arithmetic mean is the statistically best estimator of the next year's returns.

1.5.2. 基础题

Q-7. The following information is available for a portfolio:

Asset Class \ Time	Equities (60%)	Bonds (40%)
First year returns	15%	11%
Second year returns	11%	-5.6%
Third year returns	-13.86%	12%

The geometric mean return on the portfolio is closest to:

A. 3.9834%.

B. 3.5697%.

C. 4.5189%.

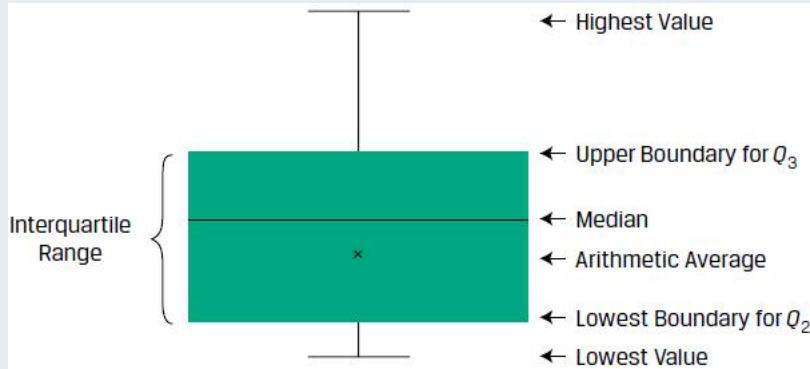### 1.6. Describe, Calculate and Interpret Quartiles, Quintiles, Deciles and Percentiles

1.6.1. 重要知识点

1.6.1.1. Quantiles

> ➢ Quartile/Quintile/Deciles/Percentile
> > ■ The third quintile: 60%, or there are three-fifths of the observations <u>fall</u>

> below that value.

> Calculation formula: $L_y = (n+1)y/100$,

   - Where L_y is the quantile position expressed in percentage.

> Box and whisker plot：

1.6.2. 基础题

Q-8. In a sample of 55 observations, if the 27th observation is ranked in ascending order, it would fall into which percentile category?

A. First quartile

B. Second quartile

C. Third quartile

1.7. Measure of Dispersion

1.7.1. 重要知识点

1.7.1.1. Absolute dispersion:

> Range = highest value-lowest value

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

> $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ (for population)

> $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ (for sample)

> Semivariance $= \dfrac{\sum_{\text{for all } X_i \le \overline{X}}(X_i - \overline{X})^2}{n-1}$

> Target Semivariance $= \dfrac{\sum_{\text{for all } X_i \le B}(X_i - B)^2}{n-1}$

> **MAD 和 variance 掌握计算和比较：**

   - Variance 比 MAD 要好，因为 variance 是连续的，处处可导。MAD 计算的是绝对值，相对比较繁琐。但是 variance 和 MAD 都是表示风险的。注意

MAD ≤ σ.

1.7.1.2. Relative dispersion

➢ **Coefficient of variation (CV)** measures the amount of risk (standard deviation) per unit of mean return.

$$CV = \frac{S_X}{\overline{X}} \times 100\%$$

1.7.1.3. The Sharpe ratio measures the reward, in terms of mean excess return, per unit of risk, as measured by standard deviation of return.

$$Sharpe\ Ratio = \frac{R_P - R_f}{\sigma_P}$$

➢ Sharpe ratio 只能用于 Rank, 没有实在含义；当小于零时，可能会得到错误的结论。

1.7.2. 基础题 2022BT.2.1.25

Q-9. An analyst gathered the following information:

Portfolio	Mean Return (%)	Sharpe ratio (%)
1	10	34
2	10	37

If the risk-free rate of return is 5.0 percent, which portfolio's coefficient of variation is larger?

A. Portfolio 1

B. Portfolio 2

C. The same

Q-10. Consider a sample of four returns for a security:

Asset	Security 1	Security 2	Security 3	Security 4
Return	3%	1%	-3%	-1%

Which of the following statements regarding the sample's mean absolute deviation (MAD) and standard deviation (SD) is true?

A. MAD ＞ SD

B. MAD ＝ SD

C. MAD ＜ SD

1.8. Skewness and Kurtosis

1.8.1. 重要知识点

1.8.1.1. Skewness 掌握概念：

> ➤ 概念：A distribution that is not symmetrical is called skewed.
> ➤ 种类：
>> ■ **Positively skewed**–A return distribution with positive skew has frequent small losses and a few extreme gains, (long right tail) (skewness > 0) (mean > median > mode);
>> ■ **Negatively skewed**–A return distribution with negative skew has frequent small gains and a few extreme losses. (long left tail) (skewness < 0) (mean < median < mode)

1.8.1.2. Kurtosis 掌握概念：

> ➤ 概念：It deals with whether or not a distribution is more or less "peaked" than a normal distribution.
> ➤ 种类：leptokurtic, normal and platykurtic

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

> ➤ 理解：
>
> A leptokurtic return distribution is more peaked and has fatter tails than the normal distribution.
> ➤ 可能在考试中会和 skew 合并考核综合知识

1.8.2. 基础题

Q-11. Which of the following is most accurate regarding a distribution of returns that has a mean greater than its median?

A. It is positively skewed.

B. It is a symmetric distribution.

C. It has positive excess kurtosis.

Q-12. A unimodal return distribution with positive excess kurtosis most likely exhibits:

A. fewer extreme outcomes than a normal distribution.

B. data that is primarily concentrated around the mean.

C. more frequent small losses and fewer extreme gains compared to a normal distribution.

Q-13. Compared to a normal distribution, a positively skewed distribution exhibits which of the following characteristics?

A. The mean is farther from the peak.

B. The mode is greater than the median, which is greater than the mean.

C. There are frequent small gains and a few extreme losses.

1.9. Expected Value and Variance

1.9.1. 重要知识点

1.9.1.1. Expected value E(X)

➢ The expected value of a random variable is the probability-weighted average of the possible outcomes of the random variable.

➢ $E(X) = \sum P_i \times X_i$

1.9.1.2. Variance Var(X) or σ²(X)

➢ The expected value (the probability-weighted average) of squared deviations from the random variable's expected value.

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

1.9.2. 基础题

Q-14. An analyst gathered the following information: the probability of economy prosperity is 80%, the probability of economy recession is 20%. For a company, when the economy is prosperity, there is 85% of probability that its EPS is $9.0 and 15% of probability that the EPS is $3.0. However, when the economy is recession, there is 10% of probability that the EPS is $9.0 and 90% of probability that the EPS is $3.0. What is the variance of this company's EPS, when the economy is prosperity?

A. 6.54

B. 4.59

C. 3.24

Q-15. A decision tree shows the potential outcomes of a company's earnings per share depending on two scenarios. Based on the provided probabilities and values, what is the expected earnings per share?

Scenario	Probabilities of scenario	Earning per share	Probabilities of earning
1	0.65	$12	0.2
		$10	0.8
2	0.35	$5	0.6
		$3	0.4

A. $6.76

B. $8.23

C. $14.60

1.10. Correlation and Covariance

1.10.1. 重要知识点

1.10.1.1. Covariance:

> Covariance is a measure of the co-movement between random variables.
> - X 与 Y 同向变化，covariance >0.
> - X 与 Y 反向变化，covariance < 0.
> - $Covariance \in (-\infty, +\infty)$.

> Covariance ranges from negative infinity to positive infinity
> - $Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$

> The covariance of a random variable with itself is its own variance.
> - $Cov(X,X) = E[(X - E(X))(X - E(X))] = \sigma^2(X)$

1.10.1.2. Correlation:

> Correlation measures the co-movement (linear association) between two random variables.

> $\rho_{XY} = \dfrac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}} = \dfrac{COV(X,Y)}{\sigma_X \sigma_Y}$

> Correlation is a number between −1 and +1.

> 理解：$\rho_{XY} \in [-1, 1]$

> - If $\rho_{x,y} = 0$, a correlation of 0 (uncorrelated variables) indicates an **absence** of any linear (straight-line) relationship between the variables.
> - Increasingly positive correlation indicates an increasingly **strong** positive linear relationship (up to 1, which indicates a perfect linear relationship).
> - Increasingly negative correlation indicates an increasingly **strong** negative (inverse) linear relationship (down to -1, which indicates a perfect inverse linear relationship).
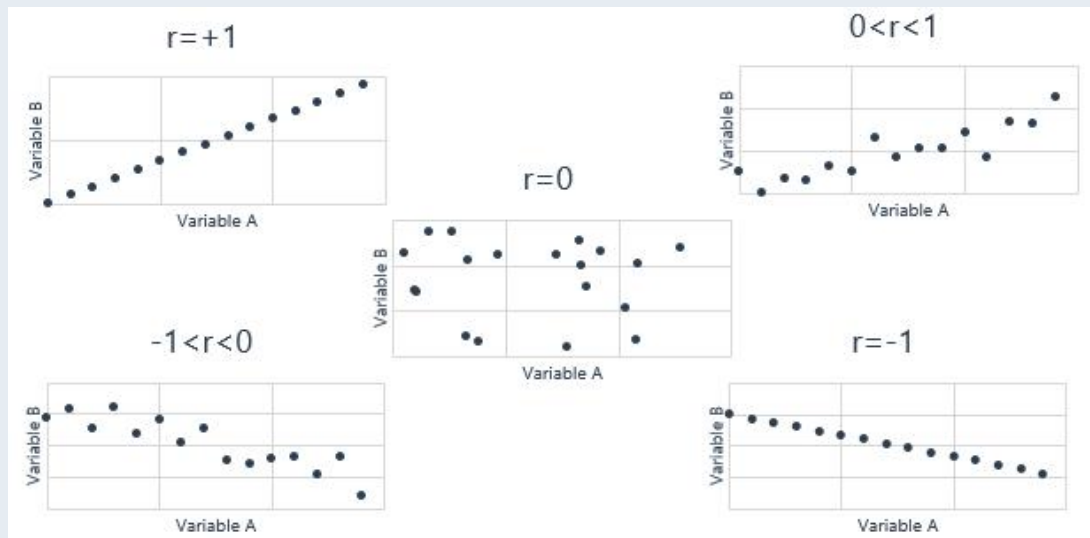
1.10.1.3. Expected return, variance and standard deviation of a portfolio

> $E(r_p) = \sum_{i=1}^{n} w_i E(R_i)$

> $\sigma^2_p = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \, cov(R_i, R_j)$

1.10.1.4. Scatter plot & limitation of correlation

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

➢ Three limitations of correlation analysis.

■ **Nonlinear relationships**: Two variables can have a strong nonlinear relation and still have a very low correlation.

■ **Outliers**: Outliers are small numbers of observations at either extreme (small or large) of a sample.

■ **Spurious correlation**: Correlations can be spurious in the sense of misleadingly pointing towards associations between variables.

1.10.2. 基础题

Q-16. The joint probability of the returns of Asset A and Asset B are given in the following figure.

Joint Probabilities	R_{B1}=0.40	R_{B2}=0.20
R_{A1}=0.20	0.30	0.20
R_{A2}=0.10	0	0.50

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

A. 0.0030.

B. 0.0024.

C. 0.6543.

Q-17. An individual wants to invest $300,000 in the following investment products:

Investment products	Expected Return	Weights	Standard deviation	Correlation
Stock	6%	80%	25%	0.2

Fund	8%	20%	30%	

What will be the rate of return and the standard deviation on the expected portfolio?

A. 6.4% and 4.84%.

B. 6.4% and 22%.

C. 6.2% and 4.84%.

Q-18.   An analyst develops the following covariance matrix of returns:

	Hedge Fund	Market Index
Hedge fund	225	90
Market index	90	64

The correlation of returns between the hedge fund and the market index is closest to:

A. 0.005.

B. 0.75

C. 0.00625.

Q-19.   Professor Irene Wang comes across the following three statements made by his students.

● Evian: Covariance lies within the range -1 < Covariance < +1

● Nicholas: The covariance of a stock with itself is equal to its own variance.

● Alex: The covariance of returns is negative when the return of one asset is above its expected value given that the return on the other asset tends to be below its expected value.

The statements made by which of the students are most likely correct?

A. Evian and Nicholas.

B. Evian and Alex.

C. Nicholas and Alex.

Q-20.   An analyst has gathered information about two securities. If the market's standard deviation of returns is 10, which of the following conclusions is most likely correct?

Security	Standard Deviation of Returns	Covariance of Returns with the Market
1	20	-160
2	11	88

A. Security 1 and Security 2 have equally strong linear relationships with the market.

B. Security 1 has a stronger linear relationship with the market than Security 2.

C. Security 2 has a stronger linear relationship with the market than Security 1.

1.11. Bayes' Formula

1.11.1. 重要知识点

1.11.1.1. Bayes' formula 掌握计算：

- **Updated probability:** Given a set of prior probabilities for an event of interest, if you receive new information, the rule for updating your probability of the event is Updated probability of event given the new information = (probability of new information given event/ unconditional probability of new information) ×prior probability of event.

  - $$P(AB) = P(B|A) \times P(A) = P(A|B) \times P(B)$$

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

- **Posterior probability（后验概率）**

  - $$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|W_1)P(W_1) + P(B|W_2)P(W_2)}$$

  - $$P(B) = P(B|W_1) \times P(W_1) + P(B|W_2) \times P(W_2) + \ldots + P(B|W_n) \times P(W_n)$$

1.11.2. 基础题

Q-21. An analyst believes Davies Company has a 40% probability of earning more than $2 per share. She estimates that the probability that Davies Company's credit rating will be upgraded is 70% if its earnings per share are greater than $2 and 20% if its earnings per share are $2 or less. Given the information that Davies Company's credit rating has been upgraded, what is the updated probability that its earnings per share are greater than $2?

A. 50%.

B. 60%.

C. 70%.

1.12. Safety First Ratio

1.12.1. 重要知识点

1.12.1.1. SFR 掌握计算及理解：

- $SFR = \left[E(R_p) - R_L \right] / \sigma_p$: the bigger, the better.

> Shortfall risk: R_L= threshold level return, minimum return required.

> Roy's safety-first criterion states that the optimal portfolio minimizes the probability that portfolio return, R_P, falls below the threshold level, R_L. In symbols, the investor's objective is to choose a portfolio that minimizes $P(R_P<R_L)$.

1.12.1.2. SFR 与 Sharpe ratio 的区别

> $$SFR = \left[E(R_p) - R_L \right] / \sigma_p$$

> $$Sharpe\ ratio = \left[E(R_P) - R_F \right] / \sigma_P$$

> Sharpe ratio will be a special case of SFR if $r_L = r_F$

1.12.2. 基础题

Q-22. On 1 January 2014, the value of an investor's portfolio is $90,000. The investor plans to donate $7,000 to charity organization and pay $3,000 to his insurance account on 31 December 2014, but meanwhile he does not want the year-end portfolio value to be below $90,000. If the expected return on the existing portfolio is 14% with a variance of 0.0225, the safety-first ratio that would be used to evaluate the portfolio based on Roy's criterion is closest to:

A. 0.193.

B. 0.465.

C. 0.415.

1.13. Normal Distribution

1.13.1. 重要知识点

1.13.1.1. Properties

> $X \sim N(\mu, \sigma^2)$

> Symmetrical distribution: skewness=0, kurtosis=3

> A linear combination of two or more normal random variables is also normally distributed.

> As the values of x gets farther from the mean, the probability density get smaller and smaller but are always positive.

1.13.1.2. Confidence intervals

> 68% confidence interval is $[\mu - \sigma, \mu + \sigma]$

> 90% confidence interval is $[\mu - 1.65\sigma, \mu + 1.65\sigma]$

> 95% confidence interval is $[\mu - 1.96\sigma, \mu + 1.96\sigma]$

> 99% confidence interval is $[\mu - 2.58\sigma, \mu + 2.58\sigma]$

1.13.1.3. Standardization

> If X ~ N(μ,σ²), then Z=(X-μ)/σ ~ N(0,1)

1.13.1.4. Cumulative probabilities for a standard normal distribution

> F(-z)=1-F(z)

> P(Z>z) = 1 −F(z)

1.13.1.5. For a group of assets, analysts can model the distribution of returns on each asset individually, or the distribution of returns on the assets as a group.

> specifies the probabilities associated with a group of random variables;

> is meaningful only when the behavior of each random variable in the group is in some way dependent upon the behavior of the others;

> apply for discrete random variables: joint probability tables;

> apply for continuous random variables: if all of the individual variables follow a normal distribution. (one of the characteristics of a normal distribution is that a linear combination of normally distributed random variables is normally distributed as well)

■ Three parameters that fully define a multivariate distribution for returns on n stocks include **mean returns** on **n** securities, **variances** of returns on **n** securities, and C_n^2=**n(n-1)/2** pairwise return **correlations**.

1.13.2. 基础题

Q-23. For a standard normal distribution, what is the probability that a random variable lies within 1 to 2 P(1< **x** <2)?

A. 13.5%

B. 27%

C. 15.5%

1.14. Lognormal Distribution

1.14.1. 重要知识点

1.14.1.1. Lognormal distribution

> Definition: If lnX is normal, then X is lognormal, which is used to describe the price of asset.

> Bounded from below by zero: the values of random variables that follow lognormal distribution are always be positive, so it is useful for modeling asset prices. $X \in (0, \infty)$

> Right skewed.

> Stock price follows lognormal distributionl, while the rate of return follows normal distribution.

1.14.2. 基础题

Q-24. In contrast to normal distributions, lognormal distributions:

A. are skewed to the left.

B. have outcomes that cannot be negative.

C. are more suitable for describing asset returns than asset prices.

Q-25. Continuous compounding refers to an interest rate that compounds continuously over time. If a stock's continuously compounded return follows a normal distribution, what is the distribution of its future price?

A. Normal

B. Uniform

C. Log-normal

1.15. Monte Carlo Smulation

1.15.1. 重要知识点

1.15.1.1. Lognormal distribution

➢ **Monte Carlo** simulation is to generate a large number of random samples from specified probability distribution(s) to represent the operation of risk in the system. It is used in planning, in financial risk management, and in valuing complex securities;

➢ Limitations:

■ The operating of Monte Carlo simulation is very complex and we must assume a parameter distribution in advance.

■ Monte Carlo simulation provides only statistical estimates, not exact results.

1.15.2. 基础题

Q-26. A Monte Carlo simulation can be used to:

A. directly provide precise valuations of call options.

B. simulate a process from historical records of returns.

C. test the sensitivity of a model to changes in assumptions.

Q-27. Which of the following statements about Monte Carlo simulation is least likely correct?

A. Monte Carlo simulation involves the use of a computer to represent the operation of a complex financial system.

B. Monte Carlo simulation requires a large number of repetitions and can effectively estimate

the future based on historical data.

C. Monte Carlo simulation is based on the repeated generation of one or more risk factors that affect security values.

1.16. Central Limit Theorem

1.16.1. 重要知识点

1.16.1.1. Central limit theorem

➢ Definition: The sampling distribution of the sample mean approaches a normal distribution as the sample size becomes large (>= 30);

➢ The mean of sample mean distribution=μ; The variance of sample mean distribution = σ^2/n.

1.16.1.2. Standard error of the sample mean

➢ Known population variance: $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

➢ Unknown population variance: $s_{\bar{x}} = s / \sqrt{n}$

1.16.2. 基础题

Q-28. A research analyst makes two statements about repeated random sampling:

Statement 1: When repeatedly drawing large samples from datasets, the sample means are approximately normally distributed.

Statement 2: The underlying population from which samples are drawn must be normally distributed in order for the sample mean to be normally distributed.

Which of the following best describes the analyst's statements?

A. Statement 1 is false; Statement 2 is true.

B. Both statements are true.

C. Statement 1 is true; Statement 2 is false.

Q-29. Which of the following statements about the Central Limit Theorem is most likely correct?

A. It states that the sample variance approaches the population variance as the sample size increases.

B. It requires that the population mean and variance are finite.

C. It assumes that the population must follow a normal distribution.

1.17. Sampling and Estimation

1.17.1. 重要知识点

1.17.1.1. Concept of sampling and estimation

- ➤ **Methods of sampling:** Probability sampling and Non-probability sampling.

    - ■ **Probability sampling:** sample random sampling, stratified random sampling , systematic sampling and cluster Sampling.

    - ■ **Non-probability sampling:** convenience Sampling and Judgmental sampling.

    - ■ **Definition of Sampling Distribution of a Statistic:** The 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.

- ➤ Sample Statistic itself is a random variable, thus following specific distribution.

- ➤ **Sampling error:** sampling error of mean=sample mean-population mean.

1.17.1.2. The desirable properties of an estimator

- ➤ **Unbiasedness:** the expected value of the estimator equals the population parameter.

- ➤ **Efficiency:** An unbiased estimator is efficient if no other unbiased estimator of the same parameter has a sampling distribution with smaller variance.

- ➤ **Consistency:** A consistent estimator is one for which the probability of estimates close to the value of the population parameter increases as sample size increases (the standard deviation of the parameter estimate decreases as the sample size increases).

    - ■ As the sample size increases, the standard error of the sample mean falls.

1.17.1.3. Point estimation: the statistic, computed from sample information, which is used to estimate the population parameter.

1.17.1.4. Interval estimation:

- ➤ Level of significance (alpha)

- ➤ Degree of Confidence (1-alpha)

- ➤ Confidence Interval = [Point Estimate +/- (reliability factor) × Standard error]

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

- ➤ **Jackknife vs. Bootstrap**

    - ■ **Jackknife produces similar results for every run.**

        - ◆ **Jackknife usually requires n repetitions. (n=sample size).**

    - ■ **Bootstrap usually gives different results because bootstrap resamples are randomly drawn.**

        - ◆ **Bootstrap needs to determine how many repetitions are**

appropriate.

1.17.1.6. Biases in sampling

- ➤ **Data snooping bias** comes from finding models by repeatedly searching through databases for patterns.
- ➤ **Sample selection bias** occurs when data availability leads to certain assets being excluded from the analysis.
  - ■ **Survivorship bias** occurs if companies are excluded from the analysis because of having gone out of business or poor performance.
  - ■ **Self-selection bias** reflects the ability of entities to decide whether or not they wish to report their attributes or results and be included in databases or samples.
  - ■ **Implicit selection bias** may exist because of a threshold enabling self-selection.
  - ■ **Backfill bias** occurs where past data, not reported or used before, is backfilled into an existing database.
- ➤ **Look-ahead bias** exists if the model uses data not available to market participants at the time when the market participants act in the model.
- ➤ **Time-period bias** is present if the time period used makes the results time-period specific or if the time period used includes a point of structural change.

1.17.2. 基础题

Q-30. An important difference between two-stage cluster sampling and stratified random sampling is that compared to stratified random sampling, two-stage cluster sampling:

A. uses all members of each sub-group (strata).

B. takes random samples all sub-groups (strata).

C. will not preserve differences in a characteristic across sub-groups.

Q-31. Which type of sampling method is being used when an individual selects a sample based on their personal preferences?

A. Simple Random Sampling.

B. Stratified Random Sampling.

C. Convenience Sampling.

Q-32. An analyst selects random subsamples from an existing dataset multiple times to estimate the standard deviation of a population. This process is called:

A. Stratified Sampling

19-51

B.   Resampling Method

C.   Systematic Sampling

1.18.   Hypothesis Testing

1.18.1.   重要知识点

1.18.1.1.   Steps of hypothesis testing

> Step 1: State null and alternative hypotheses

> Step 2: Identify the appropriate test statistic

> Step 3: Specify a level of significance

> Step 4: State a decision rule

> Step 5: Collect data and calculate the test statistic

> Step 6: Draw a conclusion

1.18.1.2.   Hypothesis testing:

> $$T\text{-}Statistic = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}; \quad T\text{-}Statistic = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

> Test Statistic follows Normal, T, Chi Square or F distributions.

> Test Statistic has formula. Calculate it with the sample data.

> This is the general formula but only for Z and T distribution.

1.18.1.3.   Relation between Confidence Intervals and Hypothesis Tests

> Confidence Interval = point estimate ± (critical value)×(standard error)

■   Center of Interval = point estimate (sample statistic)

■   Length of Interval = 2 × (critical value) × (standard error)

1.18.1.4. t-test 和 z-test 的不同应用：

Sampling from:	Statistic for small sample size (n<30)	Statistic for large sample size (n≥30)
Normal distribution with known variance	z-Statistic	z-Statistic
Normal distribution with unknown variance	t-Statistic	t-Statistic/z
Nonnormal distribution with known variance	not available	z-Statistic
Nonnormal distribution with unknown variance	not available	t-Statistic/z

1.18.1.5. Z 分布、T 分布、卡方分布、F 分布

Test type	Assumptions	H_0	Test-statistic	Critical value

	Normally distributed population, known population variance	$\mu = \mu_0$	$z = \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$	$N(0,1)$
	Normally distributed population, unknown population variance	$\mu = \mu_0$	$t = \dfrac{\overline{x} - \mu_0}{s / \sqrt{n}}$	$t(n-1)$
Mean hypothesis testing	Independent populations, unknown population variances assumed equal	$\mu_1 - \mu_2 = 0$	$t = \dfrac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$ $Where\ s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$	$t(n_1 + n_2 - 2)$
	Independent populations, unknown population variances not assumed equal	$\mu_1 - \mu_2 = 0$	$t = \dfrac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	$t*$
	Samples not independent, <u>Paired comparisons test</u> for example, two returns of stocks in the market, the return of gas and that of oil	$\mu_d = 0$	$t = \dfrac{\overline{d}}{s_{\overline{d}}}$	$t(n-1)$
		For example, the same population after and before an event, or comparison with the two different styles, which are affected by the same macro factors.		
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 population	$\sigma_1^2 = \sigma_2^2$	$F = \dfrac{s_1^2}{s_2^2}$	F(n_1-1,n_2-1)

1.18.1.6. Test Correlation

Test type	Assumptions	H_0	Test-statistic	Critical value
Correlation	Both of the variables are normally distributed	$\rho = 0$	$t = \dfrac{r - 0}{\sqrt{\dfrac{1 - r^2}{n - 2}}}$	t(n-2)

1.18.2. 基础题

Q-33. Which of the following null hypothesis has the rejection region solely in the right tail of the distribution?

A. $H_0: \mu = \mu_0$

B. $H_0: \mu \geq \mu_0$

C. $H_0: \mu \leq \mu_0$

Q-34. A sample has less than 30 data selecting from a normal distributed population with known variance. If an analyst wants to test the sample mean, which of the following distribution should be used?

A. t-student distribution

B. Z distribution

C. F distribution

Q-35. Assume that the population mean is μ, sample mean is $\bar{x}$ and $s_{\bar{x}}$ is the standard error of the sample mean. Which of the following is a condition for rejecting the null hypothesis at the 95 percent confidence interval ?

A. $(\bar{X} - \mu_0) / s_{\bar{x}} > 1.96$

B. $(\bar{X} - \mu_0) > 1.96$

C. $\mu_0 - \bar{X} / s_{\bar{x}} > 1.96$

Q-36. Investment analysts often use earnings per share (EPS) forecasts. One test of forecasting quality is the zero-mean test, which states that optimal forecasts should have a mean forecasting error of 0. (Forecasting error = Predicted value of variable −

Actual value of variable.)

Performance in Forecasting Quarterly Earnings per Share		
Number of Forecasts	Mean Forecast Error (Predicted – Actual)	Standard Deviations of Forecast Errors
100	0.06	0.20

To test whether the mean forecasting error is 0, the t-statistic calculated is most likely:

A. 3.015.

B. 3.000.

C. 0.060.

Q-37. An analyst conducts a two-tailed test to determine if earnings estimates are significantly different from reported earnings. The sample size was over 90. The computed Z-statistic is 1.30. Using a 5 percent significant level, which of the following statements is TRUE?

A. Both the null and the alternative are significant.

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

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

Q-38. A small-cap growth fund's monthly returns for the past 36 months have been consistently outperforming its benchmark. An analyst is determining whether the standard deviation of monthly returns is greater than 5%. Which of the following best describes the hypothesis to be tested?

A. $H_0: \sigma^2 \leq 0.25\%$

B. $H_a: \sigma^2 > 5\%$

C. $H_0: \sigma^2 \geq 0.25\%$

Q-39. The value of a test statistic is best determined as the difference between the sample statistic and the value of the population parameter under H_0 divided by the:

A. appropriate value from the t-distribution.

B. standard error of the sample statistic.

C. sample standard deviation.

Q-40. In setting the confidence interval for the population mean of a normal or approximately normal distribution and given that the sample size is small, Student's t-distribution is the preferred approach when the variance is:

A. large.

B. known.

C. unknown.

Q-41. Jill Batten is analyzing how the returns on the stock of Stellar Energy Corp. are related with the previous month's percent change in the US Consumer Price Index for Energy (CPIENG). Based on 248 observations, she has computed the sample correlation between the Stellar and CPIENG variables to be −0.1452. She also wants to determine whether the sample correlation is statistically significant. The critical value for the test statistic at the 0.05 level of significance is approximately 1.96. Batten should conclude that the statistical relationship between Stellar and CPIENG is:

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

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

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

Q-42. When comparing the means of two normally distributed populations, what test statistic should be used in a hypothesis test?

A. t-statistic

B. F-statistic

C. Chi-square statistic

1.19. P-Value

1.19.1. 重要知识点

1.19.1.1. P-value method

> ➢ The p-value is the smallest level of significance at which the null hypothesis can be reject

> ➢ $p-value \in [0,1]$

> ➢ p-value < α: reject H_0; p-value > α: do not reject H_0.

> ➢ P$\downarrow$, easier to reject H_0

1.19.2. 基础题

Q-43. A two-tailed test of the null hypothesis that the mean of a distribution is equal to 4.00 has a p-value of 0.0567. Using a 5% level of significance (i.e., α = 0.05), the best conclusion is to:

A. fail to reject the null hypothesis.

B. increase the level of significance to 5.67%.

C. reject the null hypothesis.

Q-44. The null hypothesis of a two-tailed test is least likely to be rejected when the p-value of the test statistic:

A. exceeds a specified level of significance.

B. falls below half of a specified level of significance.

C. falls below a specified level of significance.

1.20. Type I and Type II Errors

1.20.1. 重要知识点

1.20.1.1. Type I and type II error

> **Type I error(拒真):** reject a true null hypothesis
> - Significance level (α): the probability of making a Type I error
> - Significance level $=P(\text{Type I error})=P(H_0\times|H_0\checkmark)$

> **Type II error(取伪):** do not reject a false null hypothesis
> - $P(\text{Type II error}) = P(H_a\times|H_a\checkmark)$
> - **Power of a test**: the probability of correctly rejecting the null hypothesis when it is false.
> - Power of test $= 1 - P(\text{Type II error}) = P(H_a\checkmark|H_a\checkmark)$

1.20.2. 基础题

Q-45. What term describes the probability of correctly rejecting a false null hypothesis?

A. Significance level

B. Power of test

C. Confidence interval width

Q-46. In hypothesis testing, which of the following best describes a Type II error?

A. Rejecting a true null hypothesis

B. Rejecting a false null hypothesis

C. Failure to reject a false null hypothesis

1.21. Nonparametric tests & Tests of Independence

1.21.1. 重要知识点

1.21.1.1. 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 data are given in ranks.

■ when the hypothesis we are addressing does not concern a parameter.

1.21.1.2. 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.

➢ If test statistic χ^2 > critical value, there is sufficient evidence to conclude that size and investment type are related(not independent).

1.21.2. 基础题

Q-47. Which of the following scenarios is most suitable for using a nonparametric test?

A. When the data satisfy the assumptions of a specific distribution.

B. When the data are in the form of ranks or are measured on an ordinal scale.

C. When the population follows a lognormal distribution.

Q-48. A contingency table can be used to test:

A. a null hypothesis that rank correlations are equal to zero.

B. whether multiple characteristics of a population are independent.

C. the number of p-values from multiple tests that are less than adjusted critical values.

Q-49. A non-parametric test of independence for data in a contingency table with 6 rows and 8 columns is based on a chi-squared distribution. How many degrees of freedom does this test statistic have?

A. 14

B. 35

C. 48

Q-50. For a dataset with 50 observations, which scenario suggests parametric tests would likely perform better than non-parametric tests?

A. The data come from a skewed population.

B. The data contain ties.

C. The data follow a normal distribution.

1.22. The basics of simple linear regression model

1.22.1. 重要知识点

1.22.1.1. The basics of simple linear regression model

➢ **Function:** $Y_i = b_0 + b_1 X_i + \varepsilon_i, i = 1, ..., n$

■ **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, b0 is intercept term of the regression, b1 is slope coefficient of the regression, regression coefficient.

◆ The error term, εi is the portion of the dependent variable that is not explained by the independent variable(s) in the regression.

➢ **The assumptions of the linear regression**

■ A linear relationship exists between X and Y;

■ <u>The independent variable, X, is not random</u>, 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., $E(\varepsilon_i) = 0$)

■ The variance of the error term is constant. If not, this refers to heteroskedasticity.

■ The error term is uncorrelated across observations; ($E(\varepsilon_i\varepsilon_j) = 0$ for all $i \neq j$)

■ The error term is normally distributed.

1.22.2. 基础题

Q-51. Which of the following is least likely a necessary assumption of simple linear regression analysis?

A. The residuals are normally distributed.

B. There is a constant variance of the error term.

C. The dependent variable is uncorrelated with the residuals.

Q-52. What is the most appropriate interpretation of a slope coefficient estimate of 10.0?

A. The predicted value of the dependent variable when the independent variable is zero is 10.0.

B. For every one unit change in the independent variable, the model predicts that the dependent variable will change by 10 units.

C. For every 1-unit change in the independent variable, the model predicts that the dependent variable will change by 0.1 units.

Q-53. In the context of Ordinary Least Squares (OLS) regression, what does the intercept of the regression line represent?

A. The value of the dependent variable when the independent variable is zero.

B. The value of the independent variable when the dependent variable is zero.

C. The change in the dependent variable for a one-unit change in the independent variable.

1.23. Estimate of Regression Coefficients & Hypothesis Test

1.23.1. 重要知识点

1.23.1.1. Point estimate:

➢ $\hat{b}_1 = b_1$ $\hat{b}_0 = b_0$

➢ Calculation of $\hat{b}_1$ and $\hat{b}_0$

■ **Ordinary least squares (OLS):** Minimize the sum of squared vertical distances between the observations and the regression line (also called residuals or error terms).

◆ $\hat{b}_1 = \frac{Cov(X,Y)}{Var(X)}$; $\hat{b}_0 = \overline{Y} - \hat{b}_1\overline{X}$

➢ Regression coefficient confidence interval

■ $\hat{b}_1 \pm t_c s_{\hat{b}_1}$ ⬅ t_c 查表所得

◆ If the confidence interval at the desired level of significance does not include zero, the null is rejected, and the coefficient is said to be statistically different from zero.

1.23.1.2. Hypothesis testing about the regression coefficient

- ➤ **Significance test for a regression coefficient**
    - ■ H_0: b_1=The hypothesized value
    - ■ $t = \frac{\hat{b}_1 - hypothesized\ value\ of\ b_1}{S_{\hat{b}_1}}$, df=n-2
    - ■ Decision rule: reject H_0 if | t | > + t critical
    - ■ Rejection of the null means that the slope coefficient is <u>significantly different from zero.</u>
- ➤ **P-value Method**
    - ■ H_0: b_1= 0
    - ■ The **p-value** is the <u>smallest</u> level of significance at which the null hypothesis can be reject.
    - ■ p-value < α: reject H_0.
    - ■ reject H_0 means the coefficient is <u>significantly different from zero.</u>
- ➤ **Measure Fitness-ANOVA Table**
    - ■ **ANOVA table**

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

    - ■ **Standard error of estimate**

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

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

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

    - ■ Standard **error of estimate & coefficient of determination (R^2)**
        - ◆ The SEE is the standard deviation of the error terms in the regression.
        - ◆ The Coefficient of Determination (R2) is defined as the percentage of the total variation in the dependent variable explained by the independent variable.
        - ◆ Example: R^2 of 0.63 indicates that the variation of the independent variable explains 63% of the variation in the dependent variable.
- ➤ **Measure Fitness-F-test:** F test assesses the effectiveness of the model as a whole in explaining the dependent variable.
    - ■ Define hypothesis:

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

- ■ **F-statistic** $= F = \dfrac{MSR}{MSE} = \dfrac{RSS/k}{SSE/(n-k-1)}$

- ■ Critical value (查表)：F_α (k, n-k-1) "one-tailed" F-test; alpha=5%

- ■ Decision rule
  - ◆ Reject **H_0** : if F-statistic > F_α (k, n-k-1)

1.23.2. 基础题

Q-54. Bill Coldplay, CFA, is analyzing the performance of the Vigorous Growth Index Fund (VIGRX) over the past three years. The fund employs a passive management investment approach designed to track the performance of the MSCI US Prime Market Growth Index, a broadly diversified index of growth stocks of large U.S. companies. Coldplay estimates a regression using excess monthly returns on VIGRX (exVIGRX) as the dependent variable and excess monthly returns on the S&P 500 Index (exS&PSOO) as the independent variable. The data are expressed in decimal terms (e.g., 0.03, not 3%).

$$exVIGRX_t = b_0 + b_1(exS\&P500_t) + \varepsilon_t$$

Results from that analysis are presented in the following figures.

Critical t-values, level of significance of 0.01: One-sided, right side: +2.4411

Estimated Coefficients		
Coefficients	Coefficient Estimate	Standard Error
b_0	0.0023	0.0022
b_1	1.1163	0.0624
Partial ANOVA Table		
Source of Variation	Sum of Squares	
Regression (explained)	0.0228	
Error (unexplained)	0.0024	

Coldplay would like to test the following hypothesis: $H_0: b_1 \leq 1$ versus $H_1: b_1 > 1$ at the 1 % significance level. The calculated t-statistic and the appropriate conclusion are:

	Calculated t-statistic	Appropriate conclusion
A.	1.86	Reject H_0
B.	1.86	Fail to reject H_0
C.	2.44	Reject H_0

Q-55. Bill Coldplay, CFA, is analyzing the performance of the Vigorous Growth Index Fund (VIGRX) over the past three years. The fund employs a passive management investment approach designed to track the performance of the MSCI US Prime Market Growth

Index, a broadly diversified index of growth stocks of large U.S. companies. Coldplay estimates a regression using excess monthly returns on VIGRX (exVIGRX) as the dependent variable and excess monthly returns on the S&P 500 Index (exS&PSOO) as the independent variable. The data are expressed in decimal terms (e.g., 0.03, not 3%).

$$exVIGRX_t = b_0 + b_1(exS\&P500_t) + \varepsilon_t$$

Results from that analysis are presented in the following figures.

Estimated Coefficients		
Coefficients	Coefficient Estimate	Standard Error
b_0	0.0023	0.0022
b_1	1.1163	0.0624
Partial ANOVA Table		
Source of Variation	Sum of Squares	
Regression (explained)	0.0228	
Error (unexplained)	0.0024	

The R^2 from the regression is closest to:

A. 0.095.

B. 0.295.

C. 0.905.

Q-56. You are examining the results of a regression estimation that attempts to explain the return on equity of a company based on the variations in asset turnover. The analysis of variance output for the regression is given in the following table. The regression was based on five observations (n = 5).

ANOVA Table for Return of Equity Regressed Against Asset Turnover					
Source	df	Sum of Squares	Mean Square	F	p-value
Regression	1	95	95	31.4922	0.0112
Tesidual	3	9.1	3.0166		
Total	4	104.1			

Based on the data shown in the ANOVA table, which of the following conclusions is least accurate?

A. The sample variance of the dependent variable would be 26.025.

B. The standard error of the estimate 4.5.

C. The calculated value of the F-statistic is 31.4922. The corresponding p-value is less than 0.05, so you reject the null hypothesis of a slope equal to zero.

Q-57. From conducting a simple linear regression analysis, an analyst obtains the following

results:

Coefficient of determination (R^2)	0.9
Sum of squared regression	90
Numbers of observations	50

What is the standard error of estimate?

A.   432.07

B.   0.46

C.   0.21

1.24. Estimate of Y

1.24.1. 重要知识点

1.24.1.1. Estimate of Y

➢ Predicted values are values of the dependent variable based on the estimated regression coefficients and a prediction about the value of the independent variable.

➢ Point estimate

$$\widehat{Y} = \widehat{b}_0 + \widehat{b}_1 X$$

➢ Confidence interval estimate

$$\widehat{Y} \pm (t_c \times S_f)$$

■ t_c = the critical t-value with df=n−2

■ s_f = the standard error of the forecast

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

1.24.1.2. Forms of Simple Linear Regression

➢ Log-Lin Model

$$\ln Y = b_0 + \boldsymbol{b}_1 X$$

➢ Lin-log model

$$Y = \boldsymbol{b_0} + \boldsymbol{b_1} \ln X$$

➢ Log-log model

$$\ln Y = \boldsymbol{b_0} + \boldsymbol{b_1} \ln X$$

1.24.2. 基础题

Q-58.   Elena Vasileva recently joined EnergyInvest as a junior portfolio analyst. Vasileva's supervisor asks her to evaluate a potential investment opportunity in Amtex, a multinational oil and gas corporation based in the United States. Vasileva's supervisor

suggests using regression analysis to examine the relation between Amtex shares and returns on crude oil.

Vasileva runs a regression of Amtex share returns on crude oil returns using the monthly data she collected. Selected regression output is presented in Exhibit 1. She uses a 1% level of significance in all her tests.

Vasileva expects the crude oil return next month, Month 37, to be −0.01. She computes the standard error of the forecast to be 0.0469.

Exhibit 1	Selected Data for Crude Oil Returns and Amtex Share Returns	
	Coefficient	Standard Error
Intercept	0.0095	0.0078
Oil return	0.2354	0.0760

Critical t-values for a 1% level of significance:

One-sided, left side: −2.441

One-sided, right side: +2.441

Two-sided: ±2.728

Using information from Exhibit 2, the 99% prediction interval for Amtex share return for Month 37 is best described as:

A. $\widehat{Y}_f \pm 0.0053$

B. $\widehat{Y}_f \pm 0.0469$

C. $\widehat{Y}_f \pm 0.1279$

1.25. Tackling Big Data with Data Science

1.25.1. 重要知识点

1.25.1.1. How is fintech used in quantitative Investment analysis

➤   **Fintech:** refers to technological innovation in the design and delivery of financial services and products and it is challenging the traditional business models of incumbent financial services providers.

➤   **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**)

✓   Characteristics of Big Data

■   Volume (very large)

■   Velocity (real-time or near-real-time)

■   Variety (mainly unstructured)

✓   Type of Big Data

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

1.25.1.2. Advanced Analytical Tools

- ➤ Artificial intelligence
  - ✓ Artificial intelligence computer systems are capable of performing tasks that have traditionally required human intelligence. This is often accomplished through the use of "if then" rules
- ➤ Machine learning (ML)
  - ✓ Machine learning (ML) is a technology that has grown out of the wider AI field.
  - ✓ ML algorithms are computer programs that are able to "learn" how to complete tasks, improving their performance over time with experience.
  - ✓ Types of machine learning
    - Supervised learning: computers learn to model relationships based on labeled training data.
    - 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.

1.25.1.3. Tackling Big Data with Data Science

- ➤ Data processing methods
  - ✓ **Data capture** refers to how the data are collected and transformed into a format that can be used by the analytical process.
    - Low-latency systems (systems that operate on networks that communicate high volumes of data with minimal delay (latency)) are essential for automated trading applications that make decisions based on real-time prices and market events
    - In contrast, high-latency systems do not require access to real-time data and calculations

- ✓ **Data curation** refers to the process of ensuring data quality and accuracy through a data cleaning exercise. This process consists of reviewing all data to detect and uncover data errors (bad or inaccurate data) and making adjustments for missing data when appropriate.
- ✓ **Data storage** refers to how the data will be recorded, archived, and accessed and the underlying database design. An important consideration for data storage is whether the data are structured or unstructured and whether analytical needs require low-latency solutions.
- ✓ **Search** refers to how to query data. Big Data has created the need for advanced applications capable of examining and reviewing large quantities of data to locate requested data content
- ✓ **Transfer** refers to how the data will move from the underlying data source or storage location to the underlying analytical tool. This could be through a direct data feed, such as a stock exchange's price feed
- ➢ Text Analytics and Natural Language Processing
  - ✓ **Text analytics**
    - ■ Computer programs that analyze and derive meaning typically from large, unstructured text- or voice-based datasets, which include company filings, written reports, quarterly earnings calls, social media, email, internet postings, and surveys
  - ✓ **Natural language processing**
    - ■ Applications include translation, speech recognition, text mining, sentiment analysis, and topic analysis
    - ■ Models using NLP analysis may incorporate non-traditional information to evaluate what people are saying such as their preferences, opinions, likes, or dislikes – in an attempt to identify trends and short-term indicators about a company, a stock, or an economic event that might have a bearing on future performance

1.25.2. 基础题

Q-59. FinTech can be defined in a general way as:

A. The application of technology to traditional banking and financial activities.

B. Use of big data analytics to improve decision-making in financial markets.

C. Automation of investment processes through machine learning techniques.

Q-60. In the context of machine learning, which of the following scenarios is most indicative of an overfitted model?

A. The model treats true parameters as noise in the training data.

B. The model fails to recognize relationships within the training data.

C. The model fails to accurately predict outcomes using a dataset different from the training data.

Solutions

1. Quantitative

Q-1.　Solution: B.

The opportunity cost of not investing in a short-term corporate bonds is equal to the sum of the nominal risk-free rate of US Treasury bill and the risk premium, and the nominal risk-free rate of US Treasury bill is equal to the sum of the real risk-free rate of US Treasury bill and the inflation premium.

In this case, the real risk-free rate is 0.5%, the inflation premium is 1.5% and the default risk premium is 1%. Therefore, the nominal risk-free rate is approximately 2%, and the total opportunity cost is approximately 3%.

Q-2.　Solution: B.

A is incorrect because it takes the natural logarithm of Term Deposit 2's stated annual rate instead of its effective annual rate (EAR). Calculation: ln (1.06) = 0.058269 = 5.83%.

B is correct because the investor will be indifferent if the EAR for both term deposits is the same. Therefore, we need to find the stated annual rate with continuous compounding that corresponds to the EAR of the monthly compounded term deposit. Calculations: EAR of Term Deposit 2 = $(1 + 0.06/12)^{12} - 1 = 0.061678$. Hence, EAR of Term Deposit 1 = $0.061678 = e^r - 1$, leading to a stated annual rate for Term Deposit 1 of $r = \ln(1.061678) = 0.059851 = 5.99\%$.

C is incorrect because it is the EAR of Term Deposit 2, not the stated annual rate of Term Deposit 1. EAR of Term Deposit 2 = $(1 + 0.06/12)^{12} - 1 = 0.061678 = 6.17\%$. This answer is also closest to the calculations $e^{0.06} - 1 = 0.061837 = 6.18\%$ or $e^{6.17\%} - 1 = 0.063643 = 6.36\%$.

Q-3.　Solution: C.

The money-weighted return accounts for the money invested and provides the investor with information on the actual return he earns on his investment. The money-weighted return and its calculation are similar to the internal rate of return and a bond's yield to maturity. Amounts invested are cash outflows from the investor's perspective and amounts returned or withdrawn by the investor, or the money that remains at the end of an investment cycle, is a cash inflow for the investor.

Calculation:

Firstly, determine the timing of each cash flow;

then, using the calculator to compute IRR.

CF0=-400*100=-40,000

CF1=20,100

CF2=20,100

CPT IRR=0.33%

Q-4.   Solution: A.

To compute an exact time-weighted rate of return on a portfolio, take the following three steps:

1. Price the portfolio immediately prior to any significant addition or withdrawal of funds. Break the overall evaluation period into subperiods based on the dates of cash inflows and outflows.

2. Calculate the holding period return on the portfolio for each subperiod.

3. Link or compound holding period returns to obtain an annual rate of return for the year (the time-weighted rate of return for the year). If the investment is for more than one year, take the geometric mean of the annual returns to obtain the time-weighted rate of return over that measurement period.

$$TWRR = \sqrt[n]{\prod_{i=1}^{N}(1 + HPR_i)} - 1$$

where n=number of years;

N=number of periods.

Set 1:[(1+6%)(1-6%)(1+6%)(1-6%)]^(1/4)-1=0.9968-1=-0.0018

Set 2:[(1+11%)(1-11%)(1+11%)(1-11%)]^(1/4)-1=09939-1=-0.0061

This problem can also be analyzed based on certain properties: when the arithmetic mean is the same, the greater the dispersion (variance) of the data, the smaller the geometric mean.

We can use the sample standard deviation to help us understand the gap between the arithmetic mean and the geometric mean. The relation between the arithmetic mean ($\overline{X}$) and geometric mean (X_G) is: $X_G \approx \overline{X} - \frac{S^2}{2}$. In other words, the larger the variance of the sample, the wider the difference between the geometric mean and the arithmetic mean.

Both sets of data have an average of 0, but since the second set of cash flow has greater volatility, its geometric mean is lower.

Q-5.   Solution: A.

All else being equal, due to the different payments, PV of option 1 will be the lowest, while PV of option 3 is the highest.

Calculation:

The present value for option 1 is $24,924. PMT=-2,000, N=20, I/Y=5, CPT: PV=24,924.

The present value for option 2 is $26,171. BGN mode,PMT=-2,000, N=20, I/Y=5, CPT: PV=26,171.

The present value for option 3 is $40,000, A=2,000, discount rate=5%.

PV=A/r=2,000/0.05=40,000

Option 1 (ordinary annuity) is the last option to choose.

Q-6.   Solution: B.

Using a financial calculator, first calculate the needed funds at retirement:

N = 20, I/Y = 6, PMT =80,000, FV = 0; calculate PV =-917,593.6975.

Then use 917,593.6975 as the FV of the accumulation phase annuity:

I/Y = 6, PV = 0, PMT = -11,606.56, FV = 917,593.6975, CPT N=30.

Q-7.   Solution: C.

The portfolio return is the weighted mean return and is calculated as follows:

$$\overline{X}_1 = \sum_{i=1}^{n} w_{e1}X_{e1} + w_{b1}X_{b1} = (15\% \times 60\%) + (11\% \times 40\%) = 13.4\%$$

$$\overline{X}_2 = \sum_{i=1}^{n} w_{e2}X_{e2} + w_{b2}X_{b2} = (11\% \times 60\%) + (-5.6\% \times 40\%) = 4.36\%$$

$$\overline{X}_3 = \sum_{i=1}^{n} w_{e3}X_{e3} + w_{b3}X_{b3} = (-13.86\% \times 60\%) + (12\% \times 40\%) =- 3.52\%$$

$$Geometric\ mean = \sqrt[N]{X_1 X_2 X_3} = \sqrt[3]{(1 + 13.4\%)(1 + 4.36\%)(1 - 3.52\%)} - 1 = 0.045189$$

Q-8.   Solution: B.

B is correct because the 27th observation is located at the 48th percentile;

$$L_y = (1 + n)\left(\frac{y}{100}\right);$$

$$27 = (1 + 55) \times \left(\frac{y}{100}\right); y = \left(\frac{27 \times 100}{1 + 55}\right) = 48.21,$$

which is in the second quartile. The second quartile includes observations that are above the 25th percentile and at or below the 50th percentile.

Q-9.   Solution: A.

Sharpe ratio = [E(R$_p$)- r$_f$]/σ, based on the Sharpe ratio formula,

we can get the σ = [E(R$_p$)- r$_f$]/Sharpe ratio,

σ$_1$ = (10%-5%)/34% = 14.71%,

σ$_2$ = (10%-5%)/37% = 13.51%.

CV = σ/$\overline{X}$, CV$_1$ = 14.71%/10% = 1.471, CV$_2$ = 13.51%/10% = 1.351.

The portfolio 1's CV is larger.

Q-10.   Solution: C.

Mean Absolute Deviation (MAD) measures the average distance between each data point and the mean of the dataset, reflecting the average magnitude of deviations. In contrast, Standard Deviation (SD) measures the dispersion of the dataset by considering the squares of these deviations, which tends to give more weight to larger deviations. While all values in this sample are deviations from zero, either negative or positive, SD is likely to be higher than MAD. This is

due to the squaring of deviations in the calculation of SD, which amplifies larger deviations. In this case, with two negative deviations (-3% and -1%) and two positive deviations (1% and 3%), the squaring process in SD calculation will result in a higher value compared to MAD.

Q-11.   Solution: A.

A distribution with a mean greater than its median is positively skewed, or skewed to the right. The skew pulls the mean. Kurtosis deals with the overall shape of a distribution, not its skewness.

Q-12.   Solution: B.

Excess kurtosis thus characterizes kurtosis relative to the normal distribution. A normal distribution has excess kurtosis equal to 0. A fat-tailed distribution has excess kurtosis greater than 0, and a thin-tailed distribution has excess kurtosis less than 0. A return distribution with positive excess kurtosis—a fat-tailed return distribution—has more frequent extremely large deviations from the mean than a normal distribution.

This fat-tailed distribution is also more likely to generate observations that are near the mean, defined here as the region ±1 standard deviation around the mean.

Skewness describes the degree to which a distribution is asymmetric about its mean, not kurtosis.

Q-13.   Solution: A.

Skewness describes the degree to which a distribution is asymmetric about its mean. An asset return distribution with positive skewness has frequent small losses and a few extreme gains compared to a normal distribution. An asset return distribution with negative skewness has frequent small gains and a few extreme losses compared to a normal distribution. Zero skewness indicates a symmetric distribution of returns.

For a continuous positively skewed unimodal distribution, the mode is less than the median, which is less than the mean. For the continuous negatively skewed unimodal distribution, the mean is less than the median, which is less than the mode. For a given expected return and standard deviation, investors should be attracted by a positive skew because the mean return lies above the median. Relative to the mean return, positive skew amounts to limited, though frequent, downside returns compared with somewhat unlimited, but less frequent, upside returns.

Q-14.   Solution: B.

When the economy prosperity:

E(EPS) = 85% × 9 + 15% × 3 = 8.1

Var(EPS) = 85% × $(9-8.1)^2$ + 15% × $(3 - 8.1)^2$ = 4.59

Q-15.   Solution: B.

To find the expected earnings per share, multiply each outcome by its probability and add them up.

Under the scenario 1, the probability-weighted average is 0.2 * $12 + 0.8 * $10 = $10.4.

Under the scenario 2, the probability-weighted average is 0.6 * $5 + 0.4 * $3 = $4.2.

Expected earnings per share is 0.65 * $10.4 + 0.35 * $4.2=$8.23

Q-16.   Solution: C.

The expected returns and variance for the individual assets are determined as:

$E(R_A)=P(R_{A1},R_B)R_{A1}+P(R_{A2},R_B)R_{A2}=(0.50)(0.20)+(0.50)(0.10) =0.15$

$\sigma^2(R_A)= P(R_{A1},R_B)\{ [(R_{A1}-E(R_A)] \text{^2} + P(R_{A2},R_B)\{ [(R_{A2}-E(R_A)] \text{^2} \}$

$= (0.50)[(0.20-0.15) \text{^2}]+(0.50)[(0.10-0.15) \text{^2}] =0.0025$

$\sigma(R_A) =\sqrt{\sigma^2(R_A)} = \sqrt{0.0025} = 0.05$

$E(R_B)=P(R_{B1},R_A)R_{B1}+P(R_{B2},R_A)R_{B2}= (0.30)(0.40)+(0.70)(0.20) =0.26$

$\sigma^2(R_B)= P(R_{B1},R_A)\{ [(R_{B1}-E(R_B)] \text{^2} + P(R_{B2},R_A)\{ [(R_{B2}-E(R_B)] \text{^2} \}$

$= (0.30)[(0.40-0.26) \text{^2}]+(0.70)[(0.20-0.26) \text{^2}] =0.0084$

$\sigma(R_B) =\sqrt{\sigma^2(R_B)} = \sqrt{0.0084} \approx 0.0917$

The covariance of the asset returns is determined as:

$Cov(R_A,R_B)=P(R_{A1},R_{B1})[(R_{A1}-E(R_A)] [(R_{B1}-E(R_B)]+ P(R_{A1},R_{B2})[(R_{A1}-E(R_A)] [(R_{B2}-(R_B)]+P(R_{A2},R_{B2})[(R_{A2}-E(R_A)]$

$[(R_{B2}-E(R_B)]$

$=0.30(0.20-0.15)(0.40-0.26)+ 0.20(0.20-0.15)(0.20-0.26)+0.50(0.10-0.15)(0.20-0.26)=0.003$

The correlation between two random variables RA and RB is defined as:

$\rho(R_A,R_B) = Cov(R_A,R_B)/[\sigma(R_A) \times\sigma(R_B)] = 0.003/(0.05\times0.0917) \approx0.6543.$

Q-17.   Solution: B.

$E(X) = \sum P_i \times X_i = 80\% \times 6\% + 20\% \times 8\% = 6.4\%$

$\sigma_p^2=w_{stock}^2\sigma_{stock}^2 + w_{fund}^2\sigma_{fund}^2 + 2 \times w_{stock} \times w_{fund} \times \sigma_{stock} \times \sigma_{fund} \times \rho_{stock,fund}$

$=80\%^2 \times 25\%^2 + 20\%^2 \times 30\%^2 + 2 \times 80\% \times 20\% \times 25\% \times 30\% \times 0.2 = 4.84\%$

$\sigma_p = \sqrt{4.84\%} = 22\%$

The rate of return and the standard deviation of portfolio is 6.4% and 22%.

Q-18.   Solution: B.

The correlation between two random variables R_i and R_j is defined as $\rho(R_i, R_j) = Cov(R_i, R_j)/[\sigma(R_i)\sigma(R_j)]$. Using the subscript i to represent hedge funds and the subscript j to represent the market index, the standard deviations are $\sigma(R_i) = 225^{1/2} = 15$ and $\sigma(R_j) = 64^{1/2} = 8$. Thus, $\rho(R_i, R_j) = Cov(R_i, R_j)/[\sigma(R_i)\sigma(R_j)] = 90/(15 \times 8) = 0.75$.

Q-19.   Solution: C.

Evian is incorrect because covariance may range from negative infinity to positive infinity, correlation lies within the range of -1 and +1.

Nicholas and Alex are correct.

Q-20.   Solution: A.

The covariance measures how two variables move together. A positive covariance indicates that the two variables tend to move in the same direction, while a negative covariance suggests they move in opposite directions. However, covariance alone is difficult to interpret due to its dependence on the scale of the variables, which is why the correlation coefficient is commonly used.

The correlation coefficient is calculated as the covariance divided by the product of the standard deviations of the two variables:
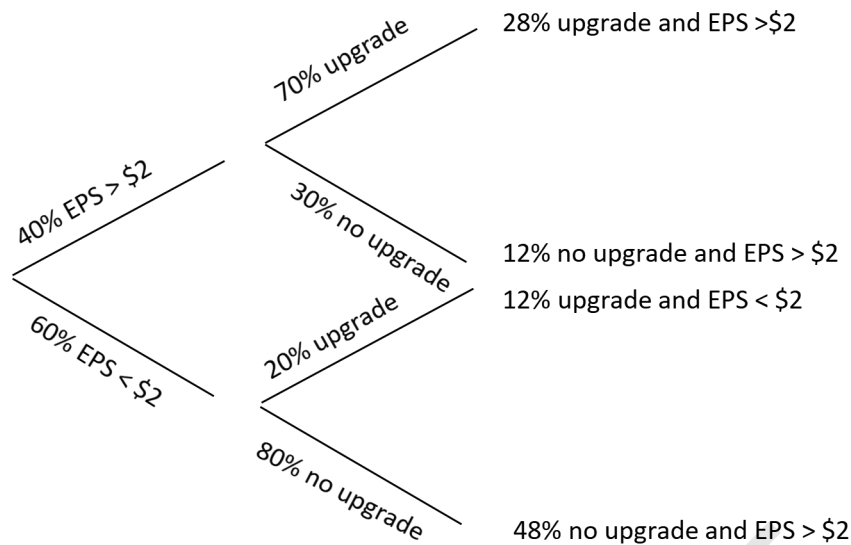
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Correlation of security 1: $-160/20^{10} = -0.8$

Correlation of security 2: $88/11*10 = 0.8$

Since the correlation of Security 1 with the market is negative (-0.8) and Security 2's correlation is positive (0.8), Security 1 and Security 2 have equally strong, but oppositely signed, linear relationships with the market.

Q-21.   Solution: C.

This is an application of Bayes formula. As the tree diagram below shows, the updated probability that earnings per share are greater than $2 is 28%/(28% + 12%)= 70%

28% upgrade and EPS >$2

70% upgrade

40% EPS > $2

30% no upgrade

12% no upgrade and EPS > $2

12% upgrade and EPS < $2

20% upgrade

60% EPS < $2

80% no upgrade

48% no upgrade and EPS > $2

Q-22.   Solution: A.

Roy's safety-first criterion states that the optimal portfolio minimizes the probability that the return of the portfolio falls below some minimum acceptable level. This minimum acceptable level is called the "threshold" level. Symbolically, Roy's safety-first criterion can be stated as:

Maximize the SFR where SFR = $[E(R_p) - R_L] / \sigma_p$

Where: R_p = portfolio return; R_L = threshold level return

R_L=10,000/90,000=11.11%, SFR = (14%-11.11%)/(0.0225$^{1/2}$) =2.89% / 0.15=19.27%

Q-23.   Solution: A.

For a standard normal distribution, the probability that a random variable lies within 1 standard of the mean is about 68%.

The probability that a random variable lies within 1.96 standard of the mean is about 95%.

The probability that a random variable lies within 1 standard deviation to 2 standard deviation is about 13.5%.

Q-24.   Solution: B.

By definition, lognormal random variables cannot have negative values.

Q-25.   Solution: C.

If a stock's continuously compounded return is normally distributed, then its future price will be log-normally distributed due to the nature of compound interest. Log-normal distributions arise naturally in finance when dealing with continuously compounded returns, since compound

growth leads to a multiplicative process, which follows a log-normal distribution. Thus, option C is the correct answer.

Option A is incorrect as normal distributions represent additive processes rather than multiplicative ones. Option B is incorrect since uniform distributions do not capture the potential positive skewness associated with stock returns.

Q-26.   Solution: C.

A characteristic feature of Monte Carlo simulation is the generation of a large number of random samples from a specified probability distribution or distributions to represent the role of risk in the system.

Q-27.   Solution: B.

While Monte Carlo simulation does require a large number of repetitions to model various outcomes, it does not necessarily provide an effective estimation of the future solely based on historical data. Monte Carlo simulation uses random sampling to simulate a range of possible outcomes based on certain input parameters, which may or may not be directly derived from historical data, and it is a complement to analytical methods. It provides only statistical estimates, not exact results.

Monte Carlo simulation indeed uses a computer to model and simulate the behavior of complex financial systems.

Monte Carlo simulation involves the repeated generation of one or more risk factors (such as interest rates, asset prices, or volatilities) to analyze their impact on security values.

Q-28.   Solution: C.

According to the central limit theorem, Statement 1 is true. Statement 2 is false because the underlying population does not need to be normally distributed in order for the sample mean to be normally distributed.

Q-29.   Solution: B.

Given a population described by any probability distribution having mean μ and finite variance σ^2, the sampling distribution of the sample mean $\overline{X}$ computed from random samples of size n from this population will be approximately normal with mean μ (the population mean) and variance σ^2/n (the population variance divided by n) when the sample size n is large.

The sample size should be large enough that the sampling distribution of the sample mean closely approximates a normal distribution. As a consequence, the CLT implies that the sample mean is a consistent estimator of the population mean as the sample size grows larger.

Q-30.   Solution: C.

With cluster sampling, the randomly selected subgroups may have different distributions of the relevant characteristic relative to the entire population. Cluster sampling uses only randomly selected subgroups, whereas stratified random sampling samples all subgroups to match the distribution of characteristics across the entire population.

Q-31.   Solution: C.

Non-probability sampling methods rely not on a fixed selection process but instead on a researcher's sample selection capabilities. Convenience Sampling, in this method, an element is selected from the population based on whether or not it is accessible to a researcher or on how easy it is for a researcher to access the element.

Probability sampling gives every member of the population an equal chance of being selected. Hence it can create a sample that is representative of the population. A simple random sample is a subset of a larger population created in such a way that each element of the population has an equal probability of being selected to the subset. The procedure of drawing a sample to satisfy the definition of a simple random sample is called simple random sampling.

In stratified random sampling, the population is divided into subpopulations (strata) based on one or more classification criteria. Simple random samples are then drawn from each stratum in sizes proportional to the relative size of each stratum in the population. These samples are then pooled to form a stratified random sample.

Q-32.   Solution: B.

Resampling involves randomly selecting different subsets of data points from an existing dataset multiple times to obtain estimates of population characteristics, such as mean and standard deviation, without making strong assumptions about underlying distributions. One common resampling technique is Bootstrap Method, which uses repeated random sampling with replacement to generate estimates of the variability associated with an estimate.

In contrast, Cluster Sampling Method involves dividing a large population into smaller groups (clusters) and selecting a sample from each group rather than choosing individual units randomly. Convenience Sampling Method selects participants according to availability, accessibility, or proximity to the researcher.

Q-33.   Solution: C.

The rejection region is solely located in the right tail of the distribution when we are testing for whether the population mean μ is greater than a specified value μ_0. Thus, H_0: $\mu \leq \mu_0$, for this test, if the sample mean is significantly greater than μ_0, we would reject the null hypothesis in favor of the alternative, which implies a rejection region in the right tail of the distribution.

μ = μ₀ represents a two-tailed test, where the rejection regions are in both tails of the distribution.

Q-34.   Solution: B.

For testing a sample mean with a small sample size and known population variance, Z distribution should be used.

Q-35.   Solution: A.

At 5% level of significance or 95% confidence interval, we calculate the z-values that correspond to 0.05/2 = 0.025 level of significance. These are +1.96 and -1.96.

We reject the null hypothesis if we find that the test statistic (($\bar{X}$ − μ₀) / s$_{\bar{x}}$) is less than -1.96 or greater than +1.96.

Q-36.   Solution: B.

The t-test is based on $t = \dfrac{X - \mu_0}{s / \sqrt{n}}$.

For this test, we have $t = \dfrac{0.06 - 0}{0.20 / \sqrt{100}} = 3$

Q-37.   Solution: C.

The sample size was over 90, which was more than 30, so z-test is appropriate. Using a 5% significant level, the critical value of a two-tailed test is 1.96. The z-statistic is 1.30, which is less than 1.96, so the analyst fail to reject null hypothesis.

Q-38.   Solution: A.

This is a one-tailed hypothesis testing with a "greater than" alternative hypothesis. A squared standard deviation is being used to obtain a test of variance.

The hypotheses are $H_0 : \sigma^2 \leq 0.25\%$ versus $H_a : \sigma^2 > 0.25\%$

Q-39.   Solution: B.

A test statistic is determined by the following formula:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Q-40.   Solution: C.

Describe the properties of Student's t-distribution and calculate and interpret its degrees of freedom.

When the sample size is small, the Student's t-distribution is preferred if the variance is unknown.

Q-41. **Solution: B.**

The calculated test statistic is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.1452\sqrt{248-2}}{\sqrt{1-(-0.1452)^2}} = -2.30177$$

Because the absolute value of t = −2.30177 is greater than 1.96, the correlation coefficient is statistically significant.

Q-42. **Solution: A.**

When conducting a hypothesis test involving the comparison of two normally distributed populations' means, the t-statistic is typically used. It relies on the differences between group means, variances, and sample sizes.

Q-43. **Solution: A.**

The p-value is the smallest level of significance at which the null hypothesis can be rejected. The smaller the p-value, the stronger the evidence against the null hypothesis. Because the p-value (0.0567) exceeds the stated level of significance (0.05), the null hypothesis cannot be rejected. A 5% confidence level does not allow the significance level to be increased beyond 5%.

Q-44. **Solution: A.**

If the p-value is less than the specified level of significance, the null hypothesis is rejected. Definetion of p-value: the smallest level of significance at which the null hypothesis can be reject.

Q-45. **Solution: B.**

The power of a statistical test measures the probability of correctly rejecting a false null hypothesis. In other words, it is the probability of obtaining statistically significant results when an alternative hypothesis is true.

Significance level (option A) is the probability of rejecting a null hypothesis when it is actually true (Type I error).

Confidence interval width (option C) refers to the range of values within which the population parameter is likely to fall with a certain degree of confidence.

Q-46. **Solution: C.**

A is incorrect because if we mistakenly reject the null hypothesis, we can only be making a Type I

error.

B is incorrect because rejecting a false null hypothesis is a correct decision and therefore not a Type II error.

C is correct because, when we make a decision in a hypothesis test, we run the risk of making either a Type I or a Type II error. These are mutually exclusive errors: If we mistakenly reject the null hypothesis, we can only be making a Type I error; if we mistakenly fail to reject the null, we can only be making a Type II error.

Q-47.　Solution: B.

Nonparametric procedures are primarily used in four situations: (1) when the data do not meet distributional assumptions, (2) when there are outliers, (3) when the data are given in ranks or use an ordinal scale, or (4) when the relevant hypotheses do not concern a parameter.

The population follows a lognormal distribution does not necessarily imply that a nonparametric test is needed. In such cases, depending on the specific analysis, parametric tests might still be suitable. Nonparametric tests are generally preferred when distributional assumptions are not met, not specifically when the distribution is lognormal.

Q-48.　Solution: B.

A contingency table is used to determine whether two characteristics of a group are independent.

Q-49.　Solution: B.

For a contingency table with 6 rows and 8 columns, the degrees of freedom would be (6-1)*(8-1) = 35.

Q-50.　Solution: C.

Nonparametric tests assume little about the underlying population distribution and therefore are generally robust against departures from normality. However, when the data comes from a normally distributed population, parametric tests such as t-tests, ANOVA, etc. are preferred since they are usually more powerful and easier to interpret compared to nonparametric tests. Outliers and ties do not necessarily rule out the possibility of performing a parametric test; however, they may lead to larger standard errors and reduced power, respectively.

In this context, the presence of outliers, ties, and skewness can still allow for the usage of non-parametric tests depending on the nature of the study. On the other hand, if the data follow a normal distribution, parametric tests are often preferred.

Q-51.　Solution: C.

The model does not assume that the dependent variable is uncorrelated with the residuals. It does assume that the independent variable is uncorrelated with the residuals.

Q-52.   Solution: B.

The slope coefficient is best interpreted as the predicted change in the dependent variable for a 1-unit change in the independent variable. If the slope coefficient estimate is 10.0 and the independent variable changes by 1 unit, the dependent variable is expected to change by 10 units. The intercept term is best interpreted as the value of the dependent variable when the independent variable is equal to zero.

Q-53.   Solution: A.

The intercept of the regression line in OLS regression represents the value of the dependent variable when the independent variable is 0. This is the point where the regression line crosses the Y-axis. It is important to note that the intercept might not always have a meaningful interpretation, especially in cases where it is unrealistic for the independent variable to be zero.

Q-54.   Solution: B.

Note that this is a one-tailed test. The critical one-tailed 1 % t-statistic with 34 degrees of freedom is approximately 2.44. The calculated t-statistic for the slope coefficient is (1.1163 - 1) / 0.0624 = 1.86. Therefore, the slope coefficient is not statistically different from one at the 1 % significance level and Coldplay should fail to reject the null hypothesis.

Q-55.   Solution: C.

SST is equal to the sum of RSS and SSE: 0.0228 + 0.0024 = 0.0252. R2 = RSS /SST= 0.0228 / 0.0252 = 0.905

Q-56.   Solution: B.

The standard error of estimate (SEE) = 3.0166^0.5 =1.7368
The sample variance of the dependent variable is the sum of squares total divided by its degrees of freedom (n − 1 = 5 − 1 = 4, as given). Thus, the sample variance of the dependent variable is 104.1 ÷ 4 = 26.025.

Q-57.   Solution: B.

A is incorrect because it is the F-statistic calculated as MSR / MSE; 90 / 0.2083 = 432.0691. For a simple linear regression, the F-Statistic is MSR divided by MSE, where the mean square regression (MSR) is the same as the sum of squares regression [SSR] and the mean square error (MSE) is calculated as the sum of squares error (SSE) adjusted by its degrees of freedom, SSE / (n

– 2); 10 / (50 − 2) = 0.2083. R^2=SSR/(SSR+SSE)=90/(90+SSE)=0.9, So, SSE is 10.

B is correct because it is the standard error of the estimate calculated as the square root of the mean square error; (0.2083)0.5 = 0.4564. The mean square error (MSE) is calculated as SSE / (n − 2); 10 / (50 − 2) = 0.2083, where SSE is calculated as R^2=SSR/(SSR+SSE), 90/(90+SSE)=0.9, So, SSE is 10.

C is incorrect because it is the mean square error (MSE) calculated as SSE / (n − 2), where SSE is the sum of squares error; 10 / (50 − 2) = 0.2083.

Q-58. Solution: C.

Confidence interval estimate: $\widehat{Y} \pm \left(t_c \times S_f\right)$

In the 1% significance level of two tailed test, the value of t statistic is 2.728. So $t_c = 2.728$. the standard error of the forecast(S_f) is 0.0469.

$t_c \times S_f = 2.728 \times 0.0469 = 0.1279$

Q-59. Solution: A.

FinTech, short for Financial Technology, refers to technological advancements in financial services and products designed to enhance user experience, reduce costs, and create new business opportunities. It encompasses various technologies like artificial intelligence, blockchain, cloud computing, mobile payments, robo-advisors, etc., used in banking and financial services. While options B and C are subsets of FinTech applications, they do not capture its full scope. Option A accurately describes FinTech's overall definition in its broadest sense.

Q-60. Solution: C.

Models that overfit the data might discover "false" relationships or "unsubstantiated" patterns that will lead to prediction errors and incorrect output forecasts. Overfitting occurs when the ML model learns the input and target dataset too precisely. In such cases, the model has been "overtrained" on the data and treats noise in the data as true parameters. An ML model that has been overfitted is not able to accurately predict outcomes using a different dataset and might be too complex.

When a model has been underfitted, the ML model treats true parameters as if they are noise and is not able to recognize relationships within the training data. In such cases, the model could be too simplistic. Underfitted models typically will fail to fully discover patterns that underlie the data.