品职教育
PZACADEMY.COM

Quantitative Methods

CFA一级知识框架图

扫码查看电子版勘误

讲师：何 旋

www.pzacademy.com

Framework

Time Value Calculation	Rates and Returns
	The Time Value of Money in Finance
Probability & Descriptive Statistics	Statistical Measures of Asset Returns
	Probability Trees and Conditional Expectations
	Portfolio Mathematics
	Simulation Methods
Inferential statistics	Estimation and Inference
	Hypothesis Testing
	Parametric and Non-Parametric Tests of Independence
Linear Regression	Simple Linear Regression
Big Data	Introduction to Big Data Techniques

Module 1

RATES AND RETURNS

Interest rate

含义	构成
① Required rate of return ② Discount rate ③ Opportunity cost	① (1 + nominal risk-free rate) = (1 + real risk-free rate)(1 + inflation premium) ② Required interest rate on a security = nominal risk-free rate + default risk premium + liquidity risk premium + maturity risk premium

Rates of Return ★★

Holding Period Return	A single specified period of time: $R = \dfrac{(P_1 - P_0) + I_1}{P_0}$ Three-year holding period return: $R = \left[(1+R_1)\times(1+R_2)\times(1+R_3)\right] - 1$
Mean Return	arithmetic mean: $\overline{R_i} = \dfrac{R_{i1} + R_{i2} + \cdots + R_{i,T-1} + R_{iT}}{T}$ geometric mean: $\overline{R_{Gi}} = \sqrt[T]{(1+R_{i1})\times(1+R_{i2})\times\cdots\times(1+R_{i,T-1})\times(1+R_{iT})} - 1$ harmonic mean: $\overline{X}_H = \dfrac{n}{\sum_{i=1}^{n}(1/X_i)}$ with $X_i > 0$ for i = 1, 2, …, n

Mean Return	结论: ① harmonic mean<= geometric mean<=arithmetic mean; $H \times A = G^2$ ② the ***more disperse*** the observations, the ***greater the difference*** between the arithmetic and geometric means. ③ ***Arithmetic mean***: ***one-period horizon;*** the ***sum of the deviations*** around the mean equals ***0;*** sensitive to extreme values, or outliers. ④ ***Geometric mean: multi-period horizon;*** backward data; represents the ***growth rate*** or ***compound rate*** of return on an investment; excellent measure of past performance ⑤ ***Harmonic mean*** : useful in ***the presence of outliers***; used most often when the ***data consist of rates and ratios***, such as P/Es.
Trimmed Mean & Winsorized Mean	结论: ① Both seek to ***minimize the impact of outliers*** in a dataset. ② Trimmed mean 是移除极值后的均值，即***截尾均值***. ③ Winsorized mean 是将极值用最近的观察值替代后的均值，即***缩尾均值***。

Money Weighted Return	定义：accounts for the money invested and provides the investor with information on the ***actual return she earns*** on her investment.
	计算: ① 找到每一期的现金流（每期的***时间间隔相同***）； ② MWRR＝计算***IRR***
	性质: 优点：衡量 the ***actual return the investor earns*** on her investment. 缺点：1）会受到现金流改变的影响，所以不能衡量基金经理的业绩；2）it ***does not allow for a return comparison*** between different individuals or different investment opportunities
Time-Weighted Returns	定义：measures the compound rate of growth of $1 initially invested in the portfolio
	计算: ① ***Price the portfolio*** immediately ***prior to any significant addition or withdrawal*** of funds； ② ***Calculate the holding period return on the portfolio for each subperiod***； ③ ***Link or compound holding period: 几何平均*** 实务: ① ***The more frequent the valuation***, the ***more accurate*** the approximation. Daily valuation is commonplace. ② Annualized time-weighted return：$R_{TW} = (1+R_1)\times(1+R_2)\times\cdots\times(1+R_{365})-1$
	性质（优点）： • ***not sensitive to the additions and withdrawals of funds***； • the TWR is the ***preferred performance measure for*** the ***evaluation of portfolios manager***

<table>
<tr><td>Annualized Return</td><td colspan="2">① $R_{annual} = (1 + R_{period})^{c} - 1$ $\longrightarrow$ $R_{weekly} = (1 + R_{daily})^{5} - 1$; $R_{weekly} = (1 + R_{annual})^{1/52} - 1$
② continuously compounded return from t to t + 1： $r_{t,t+1} = \ln(P_{t+1}/P_t) = \ln(1 + R_{t,t+1})$
③ $r_{0,T} = r_{T-1,T} + r_{T-2,T-1} + \dots + r_{0,1}$</td></tr>
<tr><td>Gross and Net Return</td><td colspan="2">• Gross return扣除Trading expense; an appropriate measure for evaluating and comparing the investment skill of asset managers。
• Net return 扣除all managerial and administrative expenses (management expenses, custodial fees, or any other administrative expenses)</td></tr>
<tr><td>Pre-Tax and After-Tax Nominal Return</td><td>$R_{After\text{-}Tax}=R_{Pre\text{-}Tax}(1\text{-}T)$</td><td rowspan="3">Gross return
Net return
Leveraged return
After-tax nominal return
After-tax real return
$\downarrow$ 计算顺序</td></tr>
<tr><td>Real Return</td><td>$(1+\text{real return}) = (1+\text{real risk-free rate})(1+\text{risk premium})$</td></tr>
<tr><td>Leveraged Return</td><td>$R_L = R_P + \frac{V_B}{V_E}(R_P - r_D)$</td></tr>
</table>

Module 2

THE TIME VALUE OF MONEY IN FINANCE

计算PV ★★

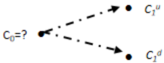
Fixed-Income Instruments	Discount	PV(Discount Bond) = $FV_t / (1 + r)^t$
	Periodic Interest	PV(Coupon Bond) = $PMT_1 / (1 + r)^1 + PMT_2 / (1 + r)^2 + \ldots + (PMT_N + FV_N) / (1 + r)^N$ PV(Perpetual Bond) = PMT / r
	Level Payments	PV(Annuity Instruments) = $A / (1 + r)^1 + A / (1 + r)^2 + \ldots + A/ (1 + r)^N$ $A = \frac{r(PV)}{1-(1+r)^{-t}}$
Equity Instruments	Constant Dividends	$PV_t = \frac{D_t}{r}$
	Constant Dividend Growth Rate	$PV_t = \frac{D_t(1+g)}{r-g} = \frac{D_{t+1}}{r-g}$
	Changing Dividend Growth Rate (两阶段模型)	$PV = \sum_{i=1}^{n} \frac{D_t(1+g_s)^i}{(1+r)^i} + \frac{E(S_{t+n})}{(1+r)^n}$ $\quad E(S_{t+n}) = \frac{D_{t+n+1}}{r-g_l}$

计算I/Y和Growth ★★

Fixed-Income Instruments	• Implied return(I/Y): discount rate or yield-to-maturity(YTM) • YTM的假设：an investor expects to receive all promised cash flows; ***reinvest*** any cash received at ***the same YTM*** • 计算：按计算器，已知PV、FV、PMT、N、求I/Y
Equity Instruments	• 计算：$r=\frac{D_{t+1}}{PV_t}+g$； $g=r-\frac{D_{t+1}}{PV_t}$ • Price-to-earnings Ratio & Forward Price-to-earnings Ratio: $\frac{PV_t}{E_t}=\frac{\frac{D_t}{E_t}(1+g)}{r-g}$ & $\frac{PV_t}{E_{t+1}}=\frac{\frac{D_{t+1}}{E_{t+1}}}{r-g}$ • 根据P/E ratio和dividend payout ratio，计算出Implied Growth，与分析师预测的Growth进行比较：Implied Growth大于预测的Growth，则判断高估；反之低估。

Cash Flow Additivity 难点

Implied Forward Rates	$F_{1,1}=(1+r_2)^2/(1+r_1)-1$
Forward Exchange Rates	汇率形式是A/B，则 $F=S_0e^{(r_A-r_B)\times t}$

<table>
<tr>
<td rowspan="2">Option Pricing
(Construct a risk-free portfolio)</td>
<td>Call Option: buy h units of the underlying asset（S_0) and sell one call option（执行价X）

$C_1^u = S_1^u - X$, $C_1^d = 0$
$V_1^u = h \times S_1^u - C_1^u$, and $V_1^d = h \times S_1^d - 0$.
$V_1^u = V_1^d$ →求出h(hedge ratio)
$V_0 = V_1^u/(1+r) = V_1^d/(1+r)$
$V_0 = h \times S_0 - c_0$→求出c_0</td>
</tr>
<tr>
<td>Put Option: buy h units of the underlying asset and buy one put option（执行价X）

$P_1^u = 0$, $P_1^d = X - S_1^d$
$V_1^u = h \times S_1^u + 0$, and $V_1^d = h \times S_1^d + P_1^d$.
$V_1^u = V_1^d$ →求出h(hedge ratio)
$V_0 = V_1^u/(1+r) = V_1^d/(1+r)$
$V_0 = h \times S_0 + P_0$→求出P_0</td>
</tr>
</table>

Module 3

STATISTICAL MEASURES OF ASSET RETURNS

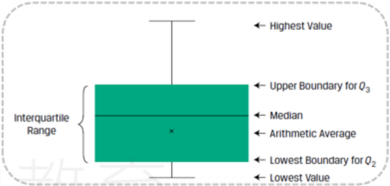
描述一组数据的基本特征

中心位置 ★★

Mean	计算：arithmetic mean 性质：the sum of the deviations around the mean equals 0 缺点：***sensitive to extreme values, or outliers.*** 对异常值的处理：①Do nothing; ②Delete all the outliers(trimmed mean); ③Replace the outliers with another value(winsorized mean)
Median	计算：Odd numbered sample: the ***(n + 1)/2 position*** Even-numbered sample: the mean of the values occupying ***the n/2 and (n + 2)/2 positions***
	优点：extreme values do not affect it; be useful ***in describing data that follow a distribution that is not symmetric*** 缺点：not use all the information about the size of the observations; focus only ***on the relative position;*** Calculating the median may also be ***more complex***.
Mode	计算: the ***most frequently occurring value*** in a distribution
	Unimodal, Bimodal and Trimodal

Quantiles

Quantiles ★★	定义：Quartile /Quintile/Deciles/Percentile
	计算：$L_y = (n+1)y/100$ → ***linear interpolation***
	性质：比如The third quartile > median
	IQR: Q_3-Q_1 ⟶ **Box and Whisker Plot**



离散程度 ★★

Range	计算：Range = maximum value – minimum value
	Advantage: ease of computation ***Disadvantage:*** It cannot tell us how the data are distributed
MAD	$MAD = \frac{\sum_{i=1}^{N}\lvert X_i - \bar{X}\rvert}{n}$

Sample Variance and Sample Standard Deviation	For population: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$ For sample: $s^2 = \dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$
	• MAD < σ • The ***gap between*** the arithmetic mean and the geometric mean: $\bar{X}_G \approx \bar{X} - \dfrac{s^2}{2}$ • ***the larger the variance*** of the sample, ***the wider the difference*** between the geometric mean and the arithmetic mean
Target Downside Deviation	$\text{Target Semivariance} = \dfrac{\sum_{\text{for all } X_i \le B}(X_i - B)^2}{n-1}$ n = the total number of observations in the sample
Coefficient of Variation	$\text{CV} = \dfrac{S_x}{\bar{X}} \times 100\%$
	• relative dispersion • ***direct comparisons of dispersion*** across ***different datasets***. • a ***scale-free*** measure (it has ***no units*** of measurement)

Skewness ★★★掌握性质

计算	$S_K = [\frac{n}{(n-1)(n-2)}]\frac{\sum_{i=1}^{n}(X_i-\overline{X})^3}{s^3} \approx (\frac{1}{n})\frac{\sum_{i=1}^{n}(X_i-\overline{X})^3}{s^3}$ a symmetrical distribution: skewness = 0	
性质	**Positive skewed**	• Mode<median<mean • right long fat tail • frequent small losses and a few extreme gains • investors should be attracted by positive skewness
	Negative skewed	• Mode>media>mean • left long fat tail • frequent small gains and a few extreme losses.

kurtosis ★★ 掌握性质

Leptokurtic	• Sample kurtosis>3； Excess kurtosis>0 • 相同σ，尖峰肥尾(more frequent extremely large deviations from the mean than a normal distribution.)
Platykurtic	Sample kurtosis<3； Excess kurtosis<0
Normal distribution	Sample kurtosis=3； Excess kurtosis=0

Covariance & Correlation ★★★

	计算	性质
Covariance	$s_{XY}=\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}$	• How two variables ***move together*** • The covariance of X with itself is equal to the variance of X • ***Positive covariance:*** the random variables vary in ***the same direction.*** • ***Negative covariance:*** the variables vary in ***the opposite direction.***
Correlation	$r_{XY}=\dfrac{s_{XY}}{s_X s_Y}$	• Correlation measures ***the strength of linear relationship*** between two random variables • ***Standardization*** of covariance, $-1 \le r_{XY} \le +1$ • If r=0, this doesn't indicates independence, it indicates an ***absence of any linear.*** • A ***positive*** correlation ***close to +1*** indicates a ***strong positive linear*** relationship; r=1 → perfect linear relationship. • A ***negative*** correlation ***close to -1*** indicates a ***strong negative*** linear relationship; r=-1 → perfect inverse linear relationship.

Limitation of Correlation Analysis

Scatter Plot: a graph that shows the relationship between the observations for two data series in two dimensions

Limitation of Correlation Analysis

- Two variables can have a ***strong nonlinear relation*** and still have a ***very low correlation.***
- Correlation also may be an ***unreliable*** measure when ***outliers*** are present.
- Correlation does ***not imply causation.***
- ***spurious correlation***
  - (1) correlation between two variables that reflects ***chance relationships*** in a particular data set
  - (2) correlation induced by a ***calculation*** that mixes each of two variables ***with a third***
  - (3) correlation between two variables arising not from a direct relation between them but ***from their relation to a third variable***.

Module 4

PROBABILITY TREES AND CONDITIONAL EXPECTATIONS

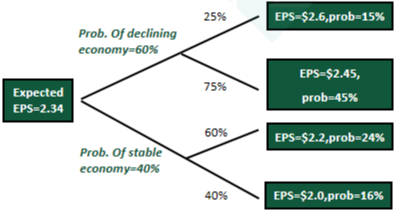
Expected Value and Variance ★

$$E(X)=P(X_1)X_1+P(X_2)X_2+\cdots+P(X_n)X_n=\sum_{i=1}^{n}P(X_i)x_i$$

$$\sigma^2(X)=E\left[X-E(X)\right]^2$$

$$\sigma^2(X)=P(X_1)[X_1-E(X)]^2+P(X_2)[X_2-E(X)]^2+\ \ldots+P(X_n)[X_n-E(X)]^2=\sum_{i=1}^{n}P(X_i)[X_i-E(X)]^2,$$

Probability Trees and Conditional Expectations ★★



会计算Conditional Expectation：

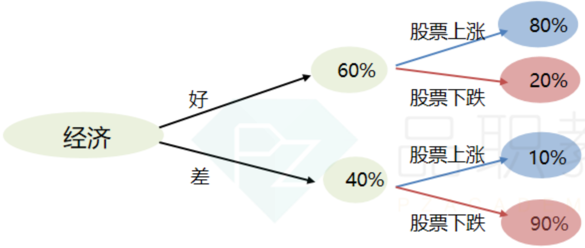
E(EPS|declining economy)

E(EPS|stable economy)

σ²(EPS|declining economy)=P(2.6|declining economy)×[2.6-E(EPS|declining economy)]²+ P(2.45|declining economy) ×[2.45-E(EPS|declining economy)]²

E(EPS)=E(EPS|declining economy)P(declining economy)+E(EPS|stable economy) P(stable economy)

Bayes' Formula ★★计算，用二叉树图形，不用记公式



注意：把非条件概率画在第一支

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

Module 5

PORTFOLIO MATHEMATICS

Portfolio Expected Return and Variance

Covariance &Correlation	$COV(X,Y) = E[(X-E(X))(Y-E(Y))]$ $\qquad \rho_{XY} = \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}}$
Independence ★★	• $P(AB)=P(A)\times P(B)$ • Independence →ρ=0; 反之不对 • ***E(XY) = E(X)E(Y) if X and Y are uncorrelated***
Portfolio Expected Return	$E(r_p) = \sum_{i=1}^{n} w_i E(R_i)$
Portfolio Variance	• $\sigma_p^2=w_1^2\sigma_1^2+w_2^2\sigma_2^2+2\rho w_1w_2\sigma_1\sigma_2$ • $\sigma_p^2=w_1^2\sigma_1^2+w_2^2\sigma_2^2++w_3^2\sigma_3^2+2\rho_{12}w_1w_2\sigma_1\sigma_2+2\rho_{13}w_1w_3\sigma_1\sigma_3+2\rho_{23}w_2w_3\sigma_2\sigma_3$

Portfolio Risk Measures

1. Safety-First Ratio	$[E(R_p)-R_L]/\sigma_p$ → Sharpe Ratio = $[E(R_p)-R_f]/\sigma_p$ Maximize SFR <=> Minimize P (R_p< R_L): shortfall risk
3. Stress testing/scenario analysis	estimating losses in extremely ***unfavorable combinations*** of ***events or scenarios***.
4. Value at risk(VaR)	***minimum value of losses*** expected over a specified time period at a given level of probability.

Module 6

SIMULATION METHODS

Lognormal Distribution and Continuous Compounding

类型	性质&计算
Lognormal distribution★★	• If lnX is normal, then X is lognormal. • Lognormal →the price of asset; normal →the return of asset • Right skewed; Bounded from below by zero (取值不能小于0) • Like the normal distribution, the lognormal distribution is completely described by two parameters(μ_L, σ_L^2).

Continuously Compounded Rates of Return ★★	• Assume that the one-period continuously compounded returns (such as $r_{0,1}$) are i.i.d(independently and identically distributed).
	  • $r_{0,T}$ is approximately normal according to the central limit theorem→P_T is lognormal
	  • $E(r_{0,T}) = E(r_{T-1,T}) + E(r_{T-2,T-1}) + . . . + E(r_{0,1}) = \mu T$
	  • $\sigma^2(r_{0,T}) = \sigma^2 T$ (平方根原则)

Monte Carlo simulation ★

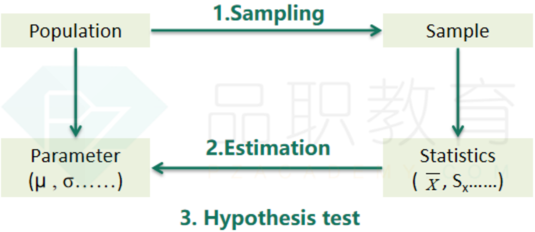
性质	
应用	• is widely used to estimate risk and return in investment applications. • can be used to ***price complex securities*** for which no analytic expression is available. • examine the ***model's sensitivity to a change in key assumptions***, such as mortgage-backed securities with complex embedded options. • address the sort of "what if " questions.
Limitations	• ***complex*** and will assume a ***parameter distribution***. • ***complement to analytical methods*** and ***provides only statistical estimates***, not exact results. (Analytical methods provide more insight into cause-and-Effect relationships.)

Module 7,8,9

难点

ESTIMATION AND INFERENCE,
HYPOTHESIS TESTING, PARAMETRIC AND NON-PARAMETRIC
TESTS OF INDEPENDENCE

Framework



Population
1.Sampling
Sample
Parameter
(μ , σ……)
2.Estimation
Statistics
($\overline{X}$, S_x……)
3. Hypothesis test

1. Sampling

Sampling methods ★	**Probability sampling**	• gives every member of the population an ***equal chance*** of being selected. • The sample is ***representative*** of the population.
		1. Simple random sampling ***2. Systematic sampling:*** select ***every kth member*** ***3. Stratified random sampling:*** draw simple random samples from each subpopulations ***in sizes proportional to the relative size of each subpopulations***. ***4. Cluster Sampling:*** the population is divided into clusters, each of which is essentially a mini-representation of the entire populations. Then ***certain clusters are chosen as a whole.*** • Disadvantage: ***lower accuracy and less representative*** • Advantage: time-efficient and cost-efficient
	non-probability sampling	• depends on other factors that is not probability considerations • May generate a ***non-representative*** sample.
		1. Convenience Sampling: an element is selected from the population ***based on whether or not it is accessible*** to a researcher 2. Judgmental Sampling: select an item ***based on a researcher's knowledge and professional judgment***. • Disadvantage: Sample selection could ***be affected by the bias of the researcher*** and might lead to ***skewed results*** • Advantage: allows researchers to go ***directly to the target population of interest***.

<table>
<tr><td rowspan="2">Sample statistic 特点</td><td colspan="2">• sampling error of the mean= sample mean- population mean
• The sample statistic itself is a random variable</td></tr>
<tr><td>Central Limit Theory
★★★</td><td>• n>=30 → sample mean ~$N(\mu, \sigma^2/n)$
• Standard error $= \sigma / \sqrt{n}$ or $s / \sqrt{n}$</td></tr>
</table>

Resampling ★	
Bootstrap	• By treating the randomly drawn sample as if it were the population, we can simulate sampling from the population by sampling from the observed sample. • the standard error of the sample mean: $S_{\overline{X}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\theta}\right)^2}$
	Advantages: • can be used to ***find the standard error or construct confidence intervals*** for the statistic of other population parameters, such as the median. • is a ***simple*** but ***powerful*** method for any complicated estimators and particularly useful ***when no analytical formula is available***. • has potential advantages in ***accuracy***.
Jackknife	• draw repeated samples while ***leaving out*** one observation at a time from the set, ***without replacing it*** • For a ***sample of size n***, jackknife usually requires ***n repetitions.***

3. Hypothesis test

步骤：检验μ ★★

<table>
<tr><td>1. 提出假设</td><td>

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

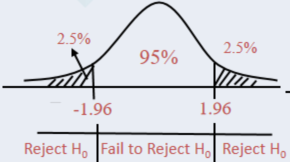
One-tailed $H_0: \mu \geq \mu_0 \ H_a: \mu < \mu_0 \ or, H_0: \mu \leq \mu_0 \ H_a: \mu > \mu_0$

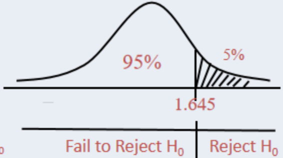
H_0 is what we want to reject

</td></tr>
<tr><td>2. 计算test statistic</td><td>

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

</td></tr>
<tr><td>3. 画分布找到critical value</td><td>





</td></tr>
<tr><td>4. 判断</td><td>

Reject H_0 if |test statistic|>critical value →***** is significantly different from ******

Fail to reject H_0 if |test statistic|<critical value →***is not significantly different from ***

</td></tr>
</table>

P – value ★★ → P – value <α → reject H_0

Type I error and Type II error ★★

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

1. Type I error ↑→ Type Ⅱ error ↓
2. Increase the Sample Size →Type I error & Type Ⅱ error ↓

Statistically and Economically Significant ★

- A strategy provides a statistically significant positive mean return. But the results may not be economically significant when we ***account for transaction costs, taxes, and risk***.
- Even if we conclude that a strategy's results are economically meaningful, the economic logic of why the strategy might work ***in the future*** should be explored before implementing it.

Multiple Tests and Interpreting Significance ★

Multiple Tests	If you run 100 tests and use a 5% significance level , you get five ***false positives***, on average.

False discovery rate (FDR)	• ***false positive result:*** Type I error
	• The false discovery approach:
	  • adjusting the p-value when you have multiple tests: rank the p-values from the various tests, from lowest to highest, and then make the following comparison: $p(\mathrm{i}) \leq \alpha \dfrac{\text{Rank of } i}{\text{Number of tests}}$
	  • Repeat the comparison: k is determined by the highest ranked p(k) for which this is a true statement

其他检验 ★★★

Test type	Assumptions	H_0	Test-statistic	Critical value
Mean hypothesis testing	Normally distributed population, known population variance	$\mu=0$	$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	N(0,1)
	Normally distributed population, unknown population variance	$\mu=0$	$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	t(n-1)
	Independent populations, unknown population variances assumed equal	$\mu_1-\mu_2=0$	$t=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}}$ ***pooled estimator*** of the common variance: $s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$t(n_1+n_2-2)$
	Independent populations, unknown population variances not assumed equal	$\mu_1-\mu_2=0$	t	t
	Samples not independent, paired comparisons test	$\mu_d=0$	$t=\bar{d}/s_{\bar{d}}$	t(n-1)
Variance hypothesis testing	Normally distributed population	$\sigma^2=\sigma_0^2$	$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2(n-1)$
	Two independent normally distributed populations	$\sigma_1^2=\sigma_2^2$	$F=s_1^2/s_2^2$	$F(n_1-1,n_2-1)$

Parameter Tests and Non-parameter Tests ★

Parameter Tests	• Rely on ***assumptions*** regarding the ***distribution*** of the population • Specific to population ***parameters***.

Non-parameter Tests	• examine quantities ***other than population parameters*** or where ***assumptions of the parameters are not satisfied***. Nonparametric tests are used:
	  • The ***assumptions*** that support a parametric test ***are not met***.
	  • When there are ***outliers***: influence the parametric statistics but not the nonparametric statistics.
	  • When ***data are ranks*** (ordinal measurement scale) rather than values.
	  • The ***hypothesis does not involve the parameters*** of the distribution, such as testing whether a variable is normally distributed.

Tests Concerning Correlation and Independence ★★

Tests Concerning Correlation	
Parametric Test of a Correlation	• H_0: ρ=0 • $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, df = n-2 • ***Two-tailed test*** • Decision rule: reject H_0 if $+t_{critical} < t$, or $t < -t_{critical}$

The Spearman Rank Correlation Coefficient	• The population ***departs from normality*** → a test based on the Spearman rank correlation coefficient, r_S.
	• Be calculated ***on the ranks of the two variables:***
	  • Rank the observations.
	  • Calculate the difference, d_i, between the ranks for each pair of observations
	  • $r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

Test type	H_0	Parameter	Test-statistic
The Spearman Rank Correlation Coefficient	$r_S = 0$	$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, df = n-2
Independence	independence	$E_{ij} = \frac{(\text{Total row } i) \times (\text{Total column } j)}{\text{Overall total}}$	$\chi^2 = \sum_{i=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, df=(r-1)(c-1)

Module 10

SIMPLE LINEAR REGRESSION

建模 → ANOVA Table分析 → 检验模型 → 预测

1.建模

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

- Y_i = ***dependent variable***, explained variable, predicted variable
- X_i = ***independent variable***, explanatory variable, predicting variable.

<table>
<tr><td rowspan="1">Assumption
★★
ε：均值为0，方差不变，不相关的正态序列</td><td>

- Linearity: The relationship between the ***Y and the X is linear.***
  - X must ***not be random (non-stochastic)***
  - The residuals are random. The residuals should not exhibit a pattern when plotted against the independent variable
- Homoskedasticity: The variance of the ***residuals is the same for all observations***
- Independence: ***The observations,*** pairs of Y_s and X_s, ***are independent of one another.*** This implies the ***regression residuals are uncorrelated*** across observations
- Normality: The residuals be ***normally distributed***
</td></tr>
<tr><td rowspan="2">Coefficient估计
★</td><td>

解释：
- ***The intercept:*** the ***value of the dependent variable*** if the value of the ***independent variable is zero***.
- ***The slope:*** the change in the Y for ***a one-unit change in the X***.
</td></tr>
<tr><td>

OLS: the sum of the squared differences between the observations on Y_i and the corresponding estimated value $\hat{Y}_i$ is minimized.

$b_1 = \frac{Cov(X,Y)}{Var(X)}$ $b_0 = \bar{Y} - b_1 \bar{X}$ **计算**
</td></tr>
</table>

2. ANOVA Table分析 ★★★

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

$$\left.\begin{array}{l} SSE=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \\ SSR=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2 \end{array}\right\} SST=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=SSR+SSE$$

Coefficient Determination (R^2)	计算: $R^2=\frac{RSS}{TSS}=1-\frac{SSE}{TSS}$
	$R^2=r_{Y\hat{Y}}^2$ (多元都成立) $R^2=r_{XY}^2$ (一元)
	解释: R^2 of 0.90 indicates that the variation of the independent variable explains 90% of the variation in the dependent variable.
SEE	计算: $SEE=\sqrt{\frac{SSE}{n-k-1}}=\sqrt{MSE}$
	性质: • The smaller the standard error, the better the fit. • The SEE is the standard deviation of the error terms in the regression.

F-test	• H_0: $b_1 = b_2 = b_3 = \ldots = b_k = 0$; H_a: at least one $b_j \neq 0$ (j = 1 to k) • $F = \dfrac{MSR}{MSE} = \dfrac{RSS/k}{SSE/(n-k-1)} \sim F(k,\ n-k-1)$ • reject H_0: if F (test-statistic) > F_c (critical value) • The F-statistic in regression analysis is ***one sided***, with the ***rejection region*** on the ***right side.***
总结: **Measures of the goodness of the fit**	• The ***coefficient of determination*** and the ***F-statistic*** are ***relative*** measures of fit • The ***standard error of the estimate*** is an ***absolute*** measure

3. 检验模型：回归分析相当于抽样估计

考试时给定条件

	Coefficient	Standard deviation	t-statistic	p-value
Intercept	$\hat{b}_0$	$s_{\hat{b}_0}$	?	0.18
Slope	$\hat{b}_1$		?	<0.001

Hypothesis Tests ★★★	• H_0: b_1=0 (没有特殊说明，题目中假设检验都是检验是否为0) • $t = \dfrac{\hat{b}_1 - B_1}{s_{\hat{b}_1}}$ • Decision rule: reject H_0 if $+t_{critical} < t$, or $t < -t_{critical}$ • Rejection of the null means that the slope coefficient is different from zero

Hypothesis Tests	
Features of simple linear regression ★★★	• the ***t-statistic*** used to test whether the ***slope coefficient is equal to zero*** and the ***t-statistic*** to test whether the ***pairwise correlation is zero*** are the same value. • the ***F-distributed test statistic*** & t-statistic used to test whether the slope coefficient is equal to zero: $t^2 = F$
Hypothesis Tests of the Intercept	$t_{intercept} = \dfrac{\hat{b}_0 - B_0}{s_{\hat{b}_0}}$

Indicator Variable or Dummy Variable ★★

定义	take on only the ***values 0 or 1*** as the ***independent variable***
解释b_0, b_1	$RET_i = b_0 + b_1 EARN_i + \varepsilon_i$ Y = monthly returns, RET over a 30-month period X = indicator variable, EARN, that takes on a value of 0 if there is no earnings announcement that month and 1 if there is an earnings announcement • ***The intercept*** (0.5629): the mean of the returns for ***non-earnings-announcement months***. • ***The slope coefficient*** (1.2098): the ***difference*** in means of returns between earnings-announcement and non-announcement months
Hypothesis Tests	• ***Test whether the mean monthly return is the same*** for both the non-earnings-announcement months and the earnings-announcement months: H_0: $\mu_{RETearnings} = \mu_{RETNon\text{-}earnings}$ • Reject H_0: there is difference in the mean RET for the earnings-announcement and non-earnings-announcements months

4.预测(Predicted Value of Y)

The predicted value for Y	给定X代入计算Y: $\hat{Y}_f = \hat{b}_0 + \hat{b}_1 X_f$ 注意：in decimal → 代入百分号; in percent → 百分号去掉
Prediction Interval ★★	$\hat{Y}_f \pm t_{\text{critical for } \alpha/2} S_f$ $\quad s_f = s_e \sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} \approx s_e$ The standard error of the forecast depends on: (***the smaller S_f*** depend on:) • ***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 closer*** the X_f is to the mean of the independent variable

Functional Forms

类型	性质
the Log-Lin model: $lnY_i = b_0 + b_1X_i$	The slope coefficient: the ***relative change in the dependent variable*** for an ***absolute change in the independent*** variable.
the Lin-Log model: $Y_i=b_0 + b_1lnX_i$	The slope coefficient: the ***absolute change in the dependent*** variable for a ***relative change in the independent*** variable.
the Log-Log model: $lnY_i=b_0 + b_1lnX_i$	This model is useful in calculating elasticities because the slope coefficient is the ***relative change in Y for a relative change in X.***

Module 11

INTRODUCTION TO BIG DATA TECHNIQUES

概念理解

Fintech

Fintech = technological innovation + financial services and products

Areas of fintech development	
Analysis of large datasets	• ***Traditional data*** • ***Alternative data from non-traditional data sources***: ***social media*** and ***sensor networks***
Analytical tools	Artificial intelligence (AI)

Big Data

1. Sources of Big Data ★★★

- *Traditional data* → *corporate data*, *financial markets*
- *Non-traditional data (alternative data)* →

Individuals	• Social media • News, reviews • Web searches, personal data	*Unstructured* Volume: *growing* dramatically
Business Processes	• Transaction data • *corporate exhaust* (corporate supply chain information, banking records, and retail point-of-sale scanner data)	*structured* data leading or real-time indicators of business performance
Sensors	• *Satellites* and Geolocation • smart phones, cameras, RFID chips • *Internet of Things*	*Unstructured* Volume: *greater*

2. Characteristics of Big Data ★★★

Volume	The *amount* of data is very *large*
Velocity	*real-time* communication
Variety	*different sources* and in a *variety* of *formats* *(Structured, Semi-structured and Unstructured)*
Veracity	Determining the *credibility* and *reliability* of different data sources is an important part.

Artificial Intelligence & Machine Learning

1. Artificial Intelligence

Artificial Intelligence: capable of performing tasks that have traditionally required human intelligence

Development of AI

- ***"Expert system"***: simulate the knowledge base and analytical abilities of human experts
- ***Neural networks***

2. Machine Learning ★★★

Definition	computer programs that are able to "learn" how to complete ***tasks***, improving their ***performance*** over time with ***experience***

Terms	• "***inputs***" (a set of variables or datasets) & "***outputs***" (the target data)
	• ***Algorithm***: "learns" from the data; ***"black box"*** approaches
	• ***Training dataset*** and ***validation dataset (evaluation dataset)***
	• ***Overfitting & underfitting***
	  • ***Overfitting:*** the ML model learns the input and target dataset ***too precisely***; be ***"over-trained"*** on the data and ***treats noise*** in the data ***as true parameters; too complex*** model; ***prediction errors*** using ***a different dataset***
	  • ***Underfitting:*** treat ***true parameters*** as if they are ***noise; too simplistic model***; ***fail to fully discover patterns***

Challenges	• Still require ***human judgement*** • The ***data*** must be ***clean*** and ***free of biases and spurious data*** • Require ***sufficiently large*** amounts of ***data***

Types of Machine Learning	
Supervised learning	based on ***labeled*** (or identified) training data • predict whether local stock market performance will be up, down, or flat
Unsupervised learning	be ***not*** given ***labeled*** data***; describe the data and their structure*** • group companies into peer groups based on their characteristics
Deep learning (AI advances)	***neural networks***, with many hidden layers, to perform multistage, non-linear data

Tackling Big Data With Data Science

computer science (including machine learning) + ***extracting information from Big Data***

Data Processing Methods
Capture: how the data are ***collected*** (***Low-latency systems***) and ***transformed into a format*** for the analytical process *Curation:* ensuring ***data quality and accuracy*** *Storage:* how the data will be ***recorded***, ***archived***, and ***accessed*** (***structured or unstructured***) *Search:* how to ***query data*** *Transfer:* move from ***the underlying data source or storage location*** to the underlying ***analytical tool***

Data Visualization →
- *Traditional structured data:* using ***tables***, ***charts***, and ***trends***
- *Non-traditional unstructured data:* 3D charts, ***heat maps***, ***tree diagrams***, and ***network graphs, "tag cloud"*** and ***"mind map"***

Applications	特点
Text Analytics	from ***large, unstructured text- or voice-based datasets;***
	Include automated information retrieval; lexical analysis, or the analysis of word frequency; identify indicators of future performance.
Natural Language Processing	the intersection of ***computer science***, ***artificial intelligence***, and ***linguistics***; ***analyze*** and ***interpret human language***.
	Include translation, speech recognition, text mining, sentiment analysis, and topic analysis

Thank You!