"Those who forget CFA Level I are condemned to repeat it"
- George Santayana

CFA Level I

Quantitative Methods

Andrew L. Berkin
Head of Research
Bridgeway Capital Management
aberkin@bridgeway.com
andrew.berkin@gmail.com

March 28, 2026

Learning Modules

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

Rates and Returns

1a

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

- Interest rate interpretation
  - Required rate of return: Minimum rate needed to make an investment
  - Discount rate: Rate needed to make an investment now give expected payout later
  - Opportunity cost: What is given up to make the investment
- These are three different ways of looking at the same thing
- $PV = FV/(1 + r)^n$ where PV = present value, FV = final value, n = # periods, r = discount rate

1a

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

- Interest rate = real risk-free rate + risk premiums
  - Inflation premium
    - Expected inflation over maturity of debt, since inflation reduces purchasing power
    - Nominal risk-free rate = real risk-free rate + inflation premium
  - Default risk premium
    - Possibility borrower may not be able to pay back on time
  - Liquidity premium
    - Loss if need to get your money quickly
  - Maturity premium
    - If market rates change over the time debt is held

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

- Holding period return (single period)
  - $R = (P_1 – P_0 + I_1) / P_0$
- Arithmetic or mean return (multiple periods)
  - $R = (R_1 + R_2 + \ldots + R_{T-1} + R_T) / T$
  - Easy, holds if identical return each period
- Geometric mean return (multiple periods)
  - $R = [(1+R_1)(1+R_2)\ldots(1+R_{T-1})(1+R_T)]^{1/T} – 1$
  - Less than arithmetic; equal only if all returns same
- Harmonic mean
  - $X_H = N / [(1/x_1)+(1/x_2) +\ldots+ (1/x_{N-1})+(1/x_N)]$; all $x>0$
  - Good if outliers, e.g. P/E ratios if some E small
- Trimmed (winsorized) means remove (replace) extremes

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

- Which to use depends on various factors
  - Are there outliers to include?
    - Yes => arithmetic mean
  - Is there compounding?
    - Yes => geometric mean
  - Are there extreme outliers?
    - Yes => harmonic, trimmed, or winsorized means

1c

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

- Money-weighted return: actual return after accounting for value & timing of cash flows
  - Closely related to internal rate of return (IRR)
  - Good for what an investor actually earned
  - Not good for comparison across different investors or investments with different cash flows
- Time-weighted return: compound growth of one unit over given time period
  - Not sensitive to cash flows
  - Preferred for evaluating portfolio managers, who don’t control cash flows

1c

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

- Money-weighted return
  - Discount rate so that sum of present value of cash flows equals zero
  - $\sum_{t=0}^{T} \frac{CF_t}{(1+IR\ )^t} = 0$
  - IRR can't be solved exactly; use Excel IRR function
  - But you can plug in given IRR to see if it works
- Time-weighted return: compound growth of one unit over given time period
  - Price portfolio, determine subperiods at times of cash flows
  - Calculate holding period return for each subperiod
  - Compound subperiod returns to get total return; get geometric mean if longer than 1 year

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

- Interest paid multiple times a year:
  - $PV = FV_N / (1+R/m)^{mN}$ where PV (FV) = present (final) value, m = # periods/year, N = # years, R = quoted (annual) rate
- To annualize return, compound by # periods in year
  - $R_{annual} = (1+R_{period})^P$ where P = # periods in a year
  - If period is longer than a year, P < 1
- Continuously compounded return r = ln(1+R)
  - Continuously compounded return from t to t+1:

    $r_{t,t+1} = \ln(P_{t+1}/P_t) = \ln(1+R_{t,t+1})$

    Example: Price at t is $100, price at t+1 is $108. Return over that period is 8% = 0.08. Continuously compounded return over that period is ln(108/100) = ln(1.08) = .07696

Calculate and interpret major return measures and describe their appropriate uses

- Gross return: Return on assets managed minus trading costs and commissions; what the manager makes, measures skill
- Net return: Gross return minus management and administrative expenses; what the investor keeps
- Pre-tax and after-tax return: Return before and after taxes; after-tax return can vary by investor tax situation
- Nominal and real return: Return before and after inflation; especially useful for comparing returns across time periods when inflation varied
- Leveraged return: Using futures or borrowed money
  - R_L = (Portfolio return) / (Portfolio equity)
    $= [R_P(V_E + V_B) - V_B r_D]/V_E = R_P + (R_P - r_D)\, V_B / V_E$

  where V_E and V_B are values of equity and debt, r_D is cost of debt.

  If $R_P < r_D$ then you lose money by leveraging

Time Value of Money in Finance

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

- Fixed income: three general patterns
  - Discount: pay initial price PV, get single cash flow FV at maturity; FV – PV = interest earned (zero-coupon bond)
    - $PV = FV_t / (1 + r)^t$
  - Periodic interest: pay PV, get periodic payments PMT and final payment + principal FV at maturity
    - $PV = PMT_1/(1+r)^1 + PMT_2/(1+r)^2 + \ldots + (PMT_N+FV_N)/(1+r)^N$
    - PV = PMT / r as N goes to infinity (perpetual bond)
  - Level payments: pay PV, get uniform payment A at periodic intervals representing interest and principal repayment
    - Mortgages and annuities
    - $A = r (PV) / [1 – (1 + r)^{-t}]$ where t = # periods

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

- Equity instruments: three general approaches
  - Constant dividends: pay initial price PV, get fixed periodic dividend D
    - $PV_t = D_t / r$ (just like perpetual bond)
  - Constant dividend growth rate: pay PV, get initial dividend D_{t+1} expected to grow at constant rate g
    - $PV_t = D_{t+1} / (r - g)$ where $r - g > 0$
  - Changing dividend growth rate: pay PV, get initial dividend D_{t+1} expected to grow at changing rate
    - $PV_t = \sum_{i=1}^{n} \frac{D_t((1+gS)^i)}{(1+r)^i} + E(S_{t+n})/(1 + r)^n$

      where terminal value $E(S_{t+n}) = D_{t+n+1} / (r - g_L)$

2b

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

- Implied return for fixed income
  - Discount bond: $r = (FV_t/PV)^{1/t} - 1$
  - Periodic interest:
    - See equation from 2a; cannot solve for r exactly
    - Can use YIELD function of Excel or Google Sheets
- Equity instruments, implied return & implied growth
  - Constant dividend growth: $r - g = D_{t+1} / PV_t$
    - Can rearrange to solve for r or g
    - In terms of P/E ratio: $PV_t / E_t = [(1+g)D_t / E_t] / (r - g)$

2c

Explain the cash flow additivity principle, its importance for the no-arbitrage condition, and its use in calculating implied forward interest rates, forward exchange rates, and option values

- Cash flow additivity principle:
  - PV of future cash flow stream = sum of PV of the cash flows
- Can use to value different cash flow streams, ensure no riskless arbitrage
- Implied forward interest rates, for example
  - Invest for time 2t at rate r_2, $FV_2 = PV(1 + r_2)^{2t}$
  - Invest for time t at rate r_1, then reinvest at rate $r_{1,1}$

    $FV_2 = PV(1 + r_1)^t (1 + r_{1,1})^t$
  - These two FV_2 must be the same, else arbitrage exists
  - If we know r_1, r_2: Can solve for $r_{1,1}$

Explain the cash flow additivity principle, its importance for the no-arbitrage condition, and its use in calculating implied forward interest rates, forward exchange rates, and option values

- Forward exchange rates example
  - 3-month t-bills pay interest at 3.61% in the USA and 2.20% in Canada. 100 USD can be converted into 132.53 CAD at the current exchange rate. What should the 3-month forward exchange rate be set at? Why?
  - Strategy 1: Invest 100 USD in a USA t-bill for 3M
    - In 3M get 100exp(.0361 × .25) = 100.91 USD
  - Strategy 2: Convert 100 USD into CAD, invest in CAN t-bill and convert back into USD in 3M by buying forward today
    - 100 USD today = 132.53 CAD
    - In 3M get 132.53 exp(.0220 × .25) = 133.26 CAD
    - Forward rate is 133.26 / 100.91 = 1.3206 (100 USD = 132.06 CAD)
  - These two forward values must be the same, else arbitrage exists

Statistical Measures of Asset Returns

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

- Measures of Central Tendency
  - Arithmetic (sample) mean: $\bar{X} = (\sum_{i=1}^{n} X_i)/n$
  - Median: Middle value (mean of middle two if even # items)
  - Mode: Most frequent value
    - Unimodal if one most frequent value
    - bimodal (multimodal) if two (more) most frequent values
- Outliers
  - Option 1: Do nothing
  - Option 2: Delete outliers, e.g. trimmed mean
  - Option 3: Replace outliers, e.g. winsorized mean
- Measures of location
  - Quartile, quintile, decile, percentile: divide into 4, 5, 10, 100
  - Interquartile range: difference between 3rd & 1st quartile

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

- Measures of dispersion
  - Range = Max – Min
  - Mean absolute deviation (MAD) = $(\sum_{i=1}^{n} |X_i - \bar{X}|)/n$
  - Variance: $s^2 = [\sum_{i=1}^{n} (X_i - \bar{X})^2]/(n - 1)$
  - Standard deviation: s (square root of s^2)
  - Downside deviation S = $\text{sqrt}([\sum_{X_i \le B}^{n} (X_i - \bar{X})^2]/(n - 1))$
    - B = target; downside deviation also called target deviation
  - Coefficient of variation CV = s / $\bar{X}$

Interpret and evaluate measures of skewness and kurtosis to address an investment problem

- Skewness: Distribution is not symmetric
  - Positive skew: many small losses, a few large gains
  - Negative skew: many small gains, a few large losses
  - Skewness $\approx [\sum_{i=1}^{n}(X_i - \bar{X})^3]/(ns^3)$
- Kurtosis: Distribution has fat or thin tails
  - Fat-tailed (leptokurtic)
  - Thin-tailed (platykurtic)
  - Roughly normal distribution (mesokurtic)
  - To get excess kurtosis K_E relative to normal distribution, subtract 3
  - $K_E \approx [\sum_{i=1}^{n}(X_i - \bar{X})^4]/(ns^4) - 3$

3d

Interpret correlation between two variables to address an investment problem

- Covariance
  - Unscaled measure of how two variables move together
  - $s_{XY} = [\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})]/(n-1)$
- Correlation
  - Scaled measure of how two variables move together
  - $r_{XY} = s_{XY} / (s_X s_Y)$
    - $-1 \leq r_{XY} \leq 1$
    - $r_{XY} = 0$ indicates no linear relationship
    - r_{XY} near 1 (-1) indicates strong positive (negative) linear relationship
  - Correlation limits
    - r_{XY} can be low but variables have strong nonlinear relationship
    - r_{XY} can be affected by outliers
    - Correlation does not imply causality; beware of spurious correlation
      - Chance, calculation mixing w/3rd variable, relation to a 3rd variable
    - Plotting X vs Y in a scatterplot can be very insightful

Probability Trees and Conditional Expectations

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

- Expected value: probability-weighted average
  - $E(X) = \sum_{i=1}^{n} P(Xi\ \ )\ X_i$
- Variance of a random variable: expected value (probability-weighted average) of squared deviations from expected value
  - $\sigma^2 = E[X - E(X)]^2 = \sum_{i=1}^{n} P(Xi\ \ )[X_i - E(X)]^2$
- Standard deviation = sqrt (variance)

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

- Conditional expected value: expected value given an event or scenario
  - $E(X \mid S) = \sum_{i=1}^{n} P(X_i \mid S)\, X_i$
- Total probability rule
  - $E(X) = E(X \mid S)P(S) + E(X \mid SC)P(S^C)$
    - S^C = "complement of S" = S does **not** occur
  - $E(X) = \sum_{i=1}^{n} E(X \mid S_i)P(S_i)$
    - If S_i mutually exclusive & exhaustive
    - Total expectation is sum of all unique (not overlapping) expectations

4b

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

- Example: Anna Liszt, CFA, follows SaltyFat, a snack food company that has provided EPS guidance of $1.50. She thinks there is a 80% chance consumer sentiment stays the same or rises, and a 20% chance it drops. If sentiment drops, she estimates a 0.30 probability of beating EPS by $0.10; otherwise, she thinks it will miss by $0.40. If sentiment stays the same or rises, she estimates a 0.75 chance of beating by $0.50 and a 0.25 chance of beating by $0.15. What is her expected value by which SaltyFat will beat or miss guidance?
  - Probability tree (kind of; see text for their version):
    - 0.80 probability sentiment stays same or rises
      - 0.75 probability => beat by $0.50 with total probability 0.60 (.80 x .75)
      - 0.25 probability => beat by $0.15 with total probability 0.20 (.80 x .25)
    - 0.20 probability sentiment falls
      - 0.30 probability => beat by $0.10 with total probability 0.06 (.20 x ..30)
      - 0.70 probability => miss by $0.40 (beat by -$0.40) with total probability 0.14 (.20 x .70)
    - Expected value of beat = E(EPS beat) = $0.28
      - 0.60($0.50) + 0.20($0.15) + 0.06($0.10) + 0.14(-$0.40) = $0.28

4c

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

- Bayes' Formula
  - $P(Event \mid Information) = \frac{P(Information \mid Event)}{P(Information)} \text{ x } P(Event)$
  - In words: updated probability of event given new info is probability of new info given the event divided by probability of the new info, multiplied by probability of event
  - Example:
    - MightyMite is rumored to be a takeover target of a larger firm, which would likely cause a jump in the stock price. You estimate the odds of an offer at 65% vs. 35% of no offer. Then MightyMite issues debt. You estimate that if there is a 20% (80%) chance they would issue debt if they were (were not) getting a takeover offer. Using Bayes' formula, what is you new estimation of the probability MightyMite gets an offer?
      - $P(Debt) = 0.20(0.65) + 0.80(0.35) = 0.41$
      - $P(Debt \mid Offer)$ = 0.20
      - $P(Offer) = 0.65$
      - $P(Offer \mid Debt)$ = (0.20 / 0.41) x 0.65 = 0.317 = 31.7%

Portfolio Mathematics

Calculate and interpret the expected value, variance, standard deviation, covariance, and correlation of portfolio returns

- Expected return of portfolio = weighted average of component returns:
  - $E(R_P) = w_1E(R_1) + w_2E(R_2) + \ldots + w_nE(R_n)$
- Variance of portfolio (measure of risk):
  - $\sigma^2(R_P) = E\{[R_P - E(R_P)]^2\}$
- Standard deviation = sqrt of variance
- Covariance (Call $E(R_i) = ER_i$):
  - $Cov(R_i,R_j) = \sigma(R_i,R_j) = \sigma_{ij} = E[(R_i - ER_i)(R_j - ER_j)]$

    $= \sum_{a=1}^{n} \sum_{b=1}^{n} P(R_{i,a}, R_{j,b})(R_{i,a} - ERi)(R_{j,b} - ER_j)$
    - Positive (negative) covariance: returns tend to move with (opposite) each other
- Correlation:
  - $\rho(R_i,R_j) = Cov(R_i,R_j) / [\sigma(R_i)\sigma(R_j)]$

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

- Example: Given the joint probability function below for the returns stock and bonds in Freedonia, calculate & interpret the covariance and correlation.

Joint Probability Function Stocks and Bonds in Freedonia			
	Rb = -4%	Rb = 5%	Rb = 10%
Rs = 20%	0.25	0	0
Rs = 8%	0	0.60	0
Rs = -10%	0	0	0.15

State of Economy	Deviations Stocks	Deviations Bonds	Product of Deviations	Prob of Condition	Prob-Wt Product
Strong	20 - 8.3	-4 - 3.5	-87.75	0.25	-21.9375
Good	8 - 8.3	5 - 3.5	-0.45	0.60	-0.27
Weak	-10 - 8.3	10 - 3.5	-118.95	0.15	-17.8425

  - $E(R_s) = 0.25(20)+0.60(8)+0.15(-10) = 8.3$
  - $E(R_b) = 0.25(-4)+0.60(5)+0.15(10) = 3.5$
  - $Cov(R_s,R_b)$ = Sum of last column = -40.05
  - $\sigma_s = Sqrt[0.25(11.7)^2+0.60(-0.3)^2+0.15(-18.3)^2] = 9.19$
  - $\sigma_b = Sqrt[0.25(-7.5)^2+0.60(1.5)^2+0.15(6.5)^2] = 4.66$
  - $\rho(R_s,R_b) = -40.05 / (9.19 \times 4.66) = -0.934$
  - Stocks and bonds are expected to be highly negatively correlated in Freedonia. Note they both have positive expected return. They provide good diversification to each other.

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

- Shortfall risk: portfolio value or return falls below some acceptable threshold over some time period
- Safety-first ratio SFRatio = $[E(R_p) - R_L] / \sigma_p$
  - R_L = threshold level
  - Safety-first optimal portfolio maximizes SFRatio
    - Assumes normal returns (ignores skew, kurtosis)
    - Can get probability from normal distribution by how many stdev the expected return $E(R_p)$ is away from limit R_L
    - If R_L is risk-free rate, SFRatio is Sharpe ratio
  - Example: Portfolio has expected return of 9% & stdev of 6%, need 3% return. SFRatio = (9 – 6)/3 = 1. Probability of not meeting that level is 15.9% (recall normal distribution has about 2/3 of its area between -1 and 1, so about 1/6 chance below -1).

Simulation Methods

6a

Explain the relationship between normal and lognormal distributions and why the lognormal distribution is used to model asset prices when using continuously compounded asset returns

- Y has lognormal distribution if ln(Y) is normal
  - Lognormal distribution bounded at 0, skewed to right
  - Prices also bounded at 0
  - Lognormal good approximation to prices
  - Normal good approximation to returns
    - Unless returns extreme; with no leverage, returns ≥ -100%
  - Like normal distribution, lognormal distribution defined by 2 parameters: mean and standard deviation
    - But they are mean and stdev of associated normal distribution
  - Continuously compounded returns from 0 to T:
    - $P_T = P_0 \exp(r_{0,T})$
    - If one-period continuously compounded returns are i.i.d. (independent and identically distributed) then returns and variance scale with time:
      - $E(r_{0,T}) = \mu T$, $\sigma^2(r_{0,T}) = \sigma^2 T$
      - This is frequently used; note stdev scales with sqrt(T), as does Sharpe ratio

6b

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

- Monte Carlo simulation: generate many potential outcomes from given probability distribution(s) to get likelihood of a range of results. Steps in implementing:
  1. Specify quantity of interest
  2. Specify time grid: range & steps, e.g. 25 years by month
  3. Specify method to generate data: distribution & transform
     - E.g. normal distribution for market return; betas to get stock returns
  4. Get simulated values
  5. Calculate quantity of interest
  6. Repeat steps 4,5 for #trials, get summary data
- For the Monte Carlo simulation itself (steps 3,4):
  1. Specify the model
  2. Specify probability distribution(s)
  3. Draw random numbers from the distribution(s)
  4. Use model to convert random #s to values of interest (e.g. prices)
- Strength: Can vary parameters
- Weaknesses: Only statistical estimates, not exact results

6c

Describe the use of bootstrap resampling in conducting a simulation based on observed data in investment applications

- Resampling: repeatedly draw samples from original observed sample to statistically infer population parameters
- Bootstrap resampling: computer simulation w/out formula
  - Monte Carlo uses formula (distribution); bootstrap uses actual data
- Steps in implementing bootstrap same as Monte Carlo
  - What differs is steps to generate the data (2nd part of previous slide)
- For the bootstrap simulation itself (steps 3,4):
  1. Use observed empirical distribution to derive properties
  2. Draw multiple values (with replacement) from empirical distribution

- Strengths: simple, good representation of actual sample
- Weakness: Only statistical estimates, not exact results

Estimation and Inference

7a

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

- Probability sampling: each member has equal chance
  - Simple random: each element has equal chance; good if homogeneous, but if not then other methods may be better
  - Stratified random: divide into subpopulations (strata), choose based on relative size of each strata; more representative, better precision
  - Cluster: divide into clusters, then use simple random to select clusters; typically lower accuracy but time & cost efficient w/large population
- Non-probability sampling: depends on factors other than probability such as judgment or convenience
  - Convenience: Select based on how easy to access; may not be representative, but quick and low cost; good for preliminary
  - Judgmental: Selectively handpick based on knowledge & judgment; can introduce bias, but quick & judgment can be important
- Sampling error: difference between observed value of a statistic & what it is intended to estimate

7b

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

- Central Limit Theorem
  - As size of random sample increases, distribution of sample means tends toward normal distribution & sampling error of sample mean is reduced
    - Sample mean $\bar{X}$ will have mean μ (population mean) & variance σ^2/n (population variance / n) when n large
  - Standard error of sample mean
    - $\sigma_{\bar{X}} = \frac{\sigma}{sqrt(n)}$ when we know population stdev σ, or
    - $s_{\bar{X}} = \frac{s}{sqrt(n)}$ when we don't know, so estimate w/sample stdev s
    - In practice, almost always use the latter
    - Sample variance: $s^2 = [\sum_{i=1}^{n}(X_i - \bar{X})^2]/(n - 1)$
    - Note standard deviation measures dispersion of data from mean; standard error measures inaccuracy of parameter estimate from sampling; a conceptual difference

7c

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

- Resampling: draw many samples from original data to get statistical inference of population parameters
  - Bootstrap: repeatedly draw samples of same size, with replacement
    - Also called model-free or non-parametric resampling, as doesn't rely on a mathematical model or distribution
    - Common: simple, powerful, estimate distribution stats & confidence intervals
    - Standard error of sample mean $s_{\bar{X}} = sqrt\{[\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\theta})^2]/(B - 1)\}$
    - B = # resamples, $\hat{\theta}_b$ = mean of a resample, $\bar{\theta}$ = mean of all $\hat{\theta}_b$
  - Jackknife: Omit one observation at a time
    - For sample of size n, need n repetitions
    - Reduce bias, find std error & confidence interval of estimators

Hypothesis Testing

8a

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

- Hypothesis testing: test a statement using sample stats. Steps:
  1. State hypothesis: null H_0 and alternative H_a
  2. Identify appropriate test statistic, e.g. t, Chi-squared, F statistics
  3. Specify level of significance, e.g. 5% or 1%; probability of false positive α
  4. State decision rule
  5. Collect data & calculate test statistic
  6. Make a decision

  - Type I error: False positive; reject true H_0
  - Type II error: False negative; fail to reject false H_0

  - Confidence level = 1 – level of significance = 1 - α
    - Higher confidence level => less chance of Type I error but more chance of Type II error
  - Power of a test: probability of correctly rejecting the null
    - Complement of Type !! error β; power of a test = 1 - β

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

- Various test statistics, what they test & distributions on p. 219
- Example: Test of a single mean
  - State hypothesis: H_0: mean $\mu = M\%$ vs H_a: mean $\mu \neq M\%$
  - Identify test statistic: t-stat with n-1 degrees of freedom
  - Specify level of significance: e.g. $\alpha = 5\%$ (two tailed)
  - State decision rule: reject H_0 if t-stat outside limit set by α
  - Calculate t-stat
  - Make a decision: For $\alpha = 5\%$, reject null if $|\text{t-stat}| > 2.069$
  - Confidence interval = $\mu \pm$ critical value $\times$ [s/sqrt(n)]

8c

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

- Parametric: either
  - concerned w/parameters e.g. mean & variance, or
  - depends on assumptions, e.g. about population distribution
- Nonparametric: Not concerned w/parameters or make minimal assumptions about population
  - Used in four situations:
    1. Distributional assumptions not satisfied, e.g. not normal
       - Often convert into ranks or test “greater than” or “less than” relationships
    2. Distribution has outliers; may be normal or not
    3. Observations are ranked
    4. Question doesn’t concern a parameter, e.g. is a sample random?

- Often do both types to see sensitivity to parameters
  - If assumptions hold, parametric preferred as more power

Parametric and Non-Parametric Tests of Independence

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

- Parametric test
  - Recall Pearson correlation: $r_{XY} = s_{XY}/(s_X s_Y)$
  - If X,Y normal, t-test to reject null H_0: ρ=0
    - $t = r \times \text{sqrt}(n - 2)/\text{sqrt}(1 - r^2)$
      - As n increases, need lower r to reject null
- Nonparametric test: Spearman rank r_S
  - Rank X, Y; break ties w/average of ranks
  - Get d_i, $\text{rank}(X_i) - \text{rank}(Y_i)$ for all i pairs
  - $r_S = 1 - 6(\sum_{b=1}^{B} d_i^2)/[n(n^2 - 1)]$
  - To test significance:
    - If n > 30) use t formula above for parametric
    - Else need special tables

Explain tests of independence based on contingency table data

- If data category or discrete, can't test independence
- Contingency table: numerical data such as counts based on categories
- Test independence w/chi-square test
  - $\chi^2 = \sum_{i=1}^{m}[(O_{ij} - E_{ij})^2/E_{ij}]$, where:
    - m = # of cells in table (group 1 count x group 2 count)
    - O_{ij} = # observations in row i and column j (frequency)
    - E_{ij} = expected # observations in row i and column j if independent
    - Test statistic has (r – 1)(c – 1) degrees of freedom; r, c = # rows, cols
  - Note: I think the formula should be:

    $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c}[(O_{ij} - E_{ij})^2/E_{ij}]$
    - Why? Consider 3x3 case. Their formula says sum to 9, but i is # rows and only goes to 3. And we never sum over j.

Explain tests of independence based on contingency table data

- Example: # USA stocks by size & B/M in Jan 2026
  - Ken French data library: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
  - Is there a relationship between size and B/M?

Count by Size and B/M				
	Low B/M	Med B/M	Hi B/M	Total
Big	411	317	108	836
Small	447	735	953	2135
Total	858	1052	1061	2971

Expected Counts Eij				
	Low B/M	Med B/M	Hi B/M	Total
Big	241.43	296.02	298.55	836
Small	616.57	755.98	762.45	2135
Total	858	1052	1061	2971

  - Example for E_{ij}: Expected Big & Low = 858 × 836 / 2971 = 241.43
  - Get $(O_{ij} - E_{ij})^2/E_{ij}$ for each cell; sum is χ^2 = 337.05

Scaled Squared Deviation (Oij - Eij)^2/Eij				
	Low B/M	Med B/M	Hi B/M	Total
Big	119.10	1.49	121.62	242.21
Small	46.64	0.58	47.62	94.84
Total	165.73	2.07	169.24	337.05

  - With (3 -1)(2 – 1) = 2 degrees of freedom, one-sided (since χ^2 can only be positive), at 5% level need χ^2 > 5.99
  - χ^2 = 337.05 > 5.99, reject null hypothesis of independence
  - Intuitively makes sense: Many more (Big, Lo) and (Small, Hi) stocks

Simple Linear Regression

10a

Describe a simple linear regression model, how the least squares criterion is used to estimate regression coefficients, and the interpretation of these coefficients

- Dependent variable Y: What is being explained
- Independent variable X: explanatory variable
  - Variations in X explain variations in Y
- Simple linear regression (SLR): estimate Y as straight line w/one independent variable X
  - $Y_i = b_0 + b_1 X_i + \varepsilon_i$, i = 1,…,n
    - Wish to minimize sum of errors squared
    - Don't know true b_0, b_1; only estimates $\hat{b}_0, \hat{b}_1$ from observations
    - Similarly, Y_i is observed, $\hat{Y}_i$ is fitted from regression parameters
  - $\hat{b}_1 = \sum_{i=1}^{n}[(Y_i - \bar{Y})(X_i - \bar{X})] / \sum_{i=1}^{n}[(X_i - \bar{X})^2$
  - $\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$
  - $\hat{b}_1$ is slope, how much $\hat{Y}$ changes when X changes
  - $\hat{b}_0$ is intercept, value of $\hat{Y}$ when X = 0

10b

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

- Four key assumptions for SLR to be valid
  1. Linearity: relation between Y and X is linear
     - Expect residuals to be random; if nonlinear may see pattern when Y or residuals plotted vs X
     - If nonlinear, can regress appropriate function f(Y) vs X
  2. Homoskedasticity: variance of residuals same for all i
     - If violated, see one set of residuals w/different magnitude; implies different regimes and thus different regressions needed
  3. Independence: observations independent
     - Residuals uncorrelated; no pattern
     - Systematic cyclic pattern could indicate dependence; may need to transform data or run separate regressions
  4. Normality: residuals normally distributed
     - Doesn't imply X and Y are normal
     - Look for presence of outliers; may need to drop/winsorize data
     - May be OK to drop assumption if sample large enough, as Central Limit Theorem implies residuals move towards linearity

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

- Goodness of fit measures
  - Coefficient of determination R^2: % of variation explained
    - $R^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 / \sum_{i=1}^{n}[(Y_i - \bar{Y})^2$; square of correlation coefficient r
    - R^2 between 0 (nothing explained) and 1 (perfect straight line)
  - Mean square regression (MSR): sum of squares regression
    - $MSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ for one dependent variable
  - Mean square error (MSE) = (sum of squares error) / (n – 2)
    - $MSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ / (n – 2) ; lower => better fit
  - F-statistic = MSR / MSE; 1-sided, higher is better

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

- Hypothesis testing of regression coefficients
  - Hypothesis tests of slope coefficient
    - Is $\hat{b}_1$ different from B_1? $t = (\hat{b}_1 - B_1) / s_{\hat{b}1}$
    - $s_{\hat{b}1}$ is std error of slope coefficient; s_e is std error of estimate (see 10d)
    - $s_{\hat{b}1} = s_e / \text{sqrt}[\sum_{i=1}^{n}[(X_i - \bar{X})^2]$
  - Test if correlation r = 0:
    - $t = r \text{ sqrt}(n - 2) / \text{sqrt}(1 - r^2)$
  - Test if positive $\hat{b}_1$ or positive r: same as above but 1-sided
  - Hypothesis tests of intercept
    - Std error of intercept: $s_{\hat{b}0} = s_e \text{ sqrt}[1/n + \bar{X}^2 / \sum_{i=1}^{n}(X_i - \bar{X})^2]$
    - $t_{intercept} = (\hat{b}_0 - B_0) / s_{\hat{b}0}$
  - Hypothesis tests of slope when X is indicator variable
    - Same as if continuous variable
  - Test of hypotheses: level of significance & p-values
    - 5% common; stricter limits => less Type 1 errors, but more Type II errors

10d

Describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret the standard error of estimate in a simple linear regression

- ANOVA: The various statistics covered in 10c
  - Example from Excel:
    - df = degrees of freedom
    - SS = Sum of Squares
    - MS = Mean Square

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regressior	5	0.073132	0.014626	36.94368	7.7724E-34
Residual	738	0.292183	0.000396		
Total	743	0.365316			

- Calculate standard error of estimate s_e from ANOVA
  - s_e = sqrt(MSE) = sqrt[$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ / (n – 2)]
  - Measure of distance between observed and predicted Y
    - Smaller s_e => better fit
  - Coefficient of determination R^2, F-statistic, s_e are measures of goodness of fit of regression line
    - R^2, F-statistic are relative
    - s_e is absolute

10e

Calculate and interpret the predicted value for the dependent variable, and a prediction interval fit for it, given an estimated linear regression model and a value for the independent variable

- Prediction using simple linear regression
  - Forecast independent variable X_f, can predict Y_f
    - $\hat{Y}_f = \hat{b}_0 + \hat{b}_1 X_f$
    - Y_f isn't exact (residuals not all 0) => need forecast uncertainty
    - Standard error of forecast $s_f = s_e \text{ sqrt}[1 + 1/n + (X_f - \bar{X})^2 / \sum_{i=1}^{n}(X_i - \bar{X})^2]$
    - What determines s_f:
      - Better the fit of regression => smaller s_e => smaller s_f
      - Larger sample size n => smaller s_f
      - Closer X_f to the mean $\bar{X}$, used in regression => smaller s_f
    - Prediction interval = $\hat{Y}_f \pm (t_{\text{critical for } \alpha/2}) \times s_f$
      - Note t-test has n-2 degrees of freedom, α is significance level e.g. 5%
    - Steps to create prediction interval:
      1. Predict Y_f given X_f
      2. Choose significance level α
      3. Determine critical value for prediction interval based on n, α
      4. Compute standard error of forecast s_f
      5. Compute (1 – α) prediction interval (e.g. 1 – 5% = 95% prediction interval
    - Prediction interval narrowest at $\bar{X}$, wider as move away

Describe different functional forms of simple linear regressions

- If relationship nonlinear, need to modify dependent or independent variable, or both
  - Example transformations: log (natural log) of dependent or independent, power law, inverse, differences. Focus on 3:
    - Log-Lin model: dependent Y logarithmic, independent X linear
    - Lin-Log model: dependent Y linear, independent X logarithmic
    - Log-Log model: both dependent Y and independent X logarithmic
  - Log-Lin: $\ln Y_i = b_0 + b_1 X_i$
    - Dep var transformed => cannot directly compare to linear (R^2, F-stat)
  - Lin-Log: $Y_i = b_0 + b_1 \ln X_i$
    - Dep var not transformed, so can directly compare to linear (R^2, F-stat)
    - If lin-log appropriate, see lower s_e, higher R^2, F-stat
  - Log-Lin: $\ln Y_i = b_0 + b_1 \ln X_i$
  - To choose correct form, examine scatterplot, residuals, goodness of fit measures (R^2, F-stat, s_e)

Introduction to Big Data Techniques

11a

Describe aspects of “fintech” that are directly relevant for the gathering and analyzing of financial data

- Fintech: technical innovations in design and delivery of financial services and products
- Directly relevant aspects include:
  - Analysis of large datasets, traditional & alternative
  - Analytical tools, such as AI

Describe Big Data, artificial intelligence, and machine learning

- Big Data: Vast amounts of data from traditional and non-traditional sources
  - Datasets typically have three (four) characteristics:
    - Volume: Very large amounts; in millions, billions, or more
    - Velocity: High speed & frequency; can be real-time
    - Variety: Many different sources & formats
    - Veracity: credibility & reliability
  - Sources: financial markets, businesses, governments, individuals, sensors, Internet of Things
  - Three main sources of alternative data:
    - Data generated by individuals
    - Data generated by business processes
    - Data generated by sensors
  - Challenges: quality, volume, appropriateness
    - Need for new techniques: AI and ML

Describe Big Data, artificial intelligence, and machine learning

- Artificial Intelligence (AI): Computer systems that typically required human intelligence
  - Early example: Expert systems, try to simulate knowledge & abilities of human experts, often with if-then rules
  - Neural networks: programming designed to simulate how brain works; useful to detect abnormal changes e.g. fraud
  - Generative AI: Growing like crazy as we speak

Describe Big Data, artificial intelligence, and machine learning

- Machine Learning (ML): Computer techniques seek to extract knowledge w/minimal assumptions
  - Algorithm "learns" by generalizing from known examples
  - Need massive amounts of data, split into three subsets
    - Training: algo finds relationships between inputs & outputs from historical data
    - Validation: verify and tune model
    - Testing: See how well it does w/new data
  - Be wary of overfitting & underfitting
  - Supervised learning: Computer learns from labeled training data, vs unsupervised learning
  - Deep learning: Use neural nets, often w/many hidden layers to identify patterns; can be supervised or unsupervised

Describe applications of Big Data and Data Science to investment management

- Data science: combine computer science, statistics, other fields to extract information from data
  - Data processing methods: Capture, curation (cleaning), storage, search, transfer
  - Data visualization: Important for understanding data
  - Text analytics: analyze & derive meaning from written or spoken words
  - Natural language processing (NLP): combining computer science, AI, linguistics to analyze & interpret human language
    - Uses include translation, speech recognition, text mining, sentiment analysis, topic analysis