



## 2024 Level 1 - Quantitative Methods

Readings	Page
Rates and Returns	2
The Time Value of Money in Finance	8
Statistical Measures of Asset Returns	17
Probability Trees and Conditional Expectations	23
Portfolio Mathematics	27
Simulation Methods	31
Estimation and Inference	36
Hypothesis Testing	41
Parametric and Non-Parametric Tests of Independence	45
Simple Linear Regression	49
Introduction to Big Data Techniques	58

This document should be used in conjunction with the corresponding readings in the 2024 Level 1 CFA® Program curriculum. Some of the graphs, charts, tables, examples, and figures are copyright 2023, CFA Institute. Reproduced and republished with permission from CFA Institute. All rights reserved.

Required disclaimer: CFA Institute does not endorse, promote, or warrant accuracy or quality of the products or services offered by MarkMeldrum.com. CFA Institute, CFA®, and Chartered Financial Analyst® are trademarks owned by CFA Institute.

## **Rates and Returns**

- a. 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 distinct types of risk
- b. calculate and interpret different approaches to return measurement over time and describe their appropriate uses
- c. compare the money-weighted and time-weighted rates of return and evaluate the performance of portfolios based on these measures
- d. calculate and interpret annualized return measures and continuously compounded returns, and describe their appropriate uses
- e. calculate and interpret major return measures and describe their appropriate uses

## Rates and Returns

Page 1

**Interest rates (r) - can be thought of in 3 ways**

**1/ required rates of return** → determining the FV of a PV

**2/ discount rate** → determining the PV of a FV

**3/ opportunity cost** → value forgone (current consumption vs. saving)

### Determinants of Interest Rates

vary over time and continuously change

$r = r_f$  → real default risk-free rate (single period)

- + inflation premium → expected inflation over a period of time
- + default risk premium → compensates for credit risk
- + liquidity premium → risk of loss vs. fair value if an investment needs to be converted to cash quickly
- + maturity premium → compensation for greater price sensitivity from changes in rates

Page 2

• **nominal risk-free rate:**

$$(1 + r) = (1 + r_f)(1 + \pi^e) \quad \text{or} \quad r = r_f + \pi^e$$

Note: all rates are quoted on an annual basis

e.g. 3-mos. T-Bill @ 4% is  $.04/4 = 1\%$  over 3 months

### Rates of Return/

#### 1/ HPR - holding period return

$$R = \frac{(P_t - P_0) + I}{P_0}$$

e.g./  
 $P_t$  - ending price 105  
 $P_0$  - beginning price 100  
 $I$  - all income 0

$$HPR = \frac{(105 - 100)}{100} = 5\%$$

#### - multi-period HPR

$$R = [(1 + HPR_1)(1 + HPR_2)(1 + HPR_3)] - 1$$

#### 2/ Arithmetic or Mean Return

$$\bar{R}_i = \frac{1}{T} \cdot \sum_{t=1}^T R_{it}$$

#### 3/ Geometric Mean Return

$$\bar{R}_{Gi} = [(1 + R_{i1})(1 + R_{i2}) \dots (1 + R_{iT})]^{1/T} - 1$$

b

e.g./  $R_{i1} = -50\%$   
 $R_{i2} = 35\%$   
 $R_{i3} = 27\%$

$$[(.5)(1.35)(1.27)]^{1/3} - 1 = -5\% \quad \text{- the growth rate or compounded return on an investment}$$

### Rates of Return/

$$R_G \leq R_A$$

growth of \$1

- unless all observations are equal, then  $R_G = R_A$   
 - as the variability in the data increases, difference between  $R_G$  and  $R_A$  increases

$PV(1 + R_G)^N = FV$   
 - use to estimate  $E(R)$  over multiple periods

avg. return over time  $PV(1 + R_A)^N \neq FV$   
 - use to estimate  $E(R)$  over one period

#### 4/ Harmonic Mean

$$\bar{X}_H = \frac{n}{\sum(1/X)} \quad \text{- arithmetic mean - all obs. have equal weight}$$

→ obs. weight inversely proportional to its magnitude  
 ∴ reduces the effect of outliers

b

- most often used with ratios (amount/unit)

e.g./ P/E 45, 15, 15  
 $R_A = 25$

$$\bar{X}_H = \frac{3}{\sum(1/45 + 1/15 + 1/15)} = 3 / .15 = 19.28$$

### Rates of Return/

#### 4/ Harmonic Mean

- applied e.g. - dollar-cost averaging (typical DC strategy)

1,000/month for 2 months in a stock  $P_0 = 10$   $P_1 = 15$

$$\bar{X}_H = \frac{2}{\left(\frac{1}{10} + \frac{1}{15}\right)} = 2 / .1\dot{6} = 12/\text{sh.}$$

Proof:  $1000/10 = 100$  sh.  $1000/15 = 66.67$  sh.

$$2000 / 166.67 \text{ sh.} = 12/\text{sh.}$$

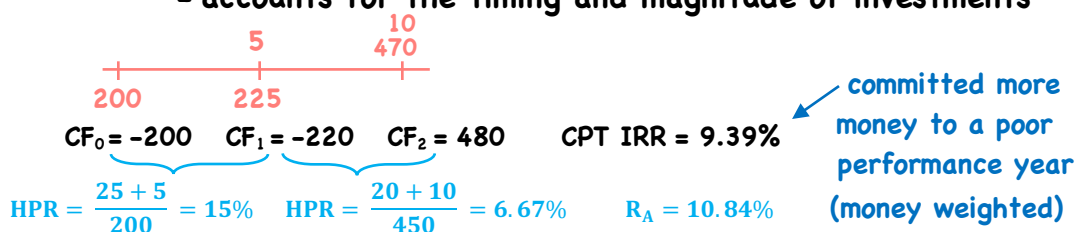
Relationship:  $R_A \times \bar{X}_H = R_G^2$

#### 5/ Others

- b
- removes outliers (on both sides)
- a) trimmed mean (common with CPI)
    - remove a %'age from both the largest and smallest
    - e.g. 100 obs., 8% trimmed = 84 obs. (8 highest, 8 lowest)
  - b) winsorized mean - replacing values at both end with the cutoff value
    - e.g. 100 obs. obs. 1-8 replaced all = obs. 9
    - obs. 93-100 replaced all = obs. 92

### • Money-weighted Return (IRR, YTM)

- accounts for the timing and magnitude of investments



- not comparable across investors/investments
- mwrr represents what 'your' money earned, not what \$1 could earn

### • Time-weighted returns (= $R_G$ )

- the growth of \$1 over a given time period
- comparable across investments

- c
- break investment period into holding periods (determined by any significant cash in/out-flows)
  - calculate each HPR, then compound the HPRs, express annually

• Time-weighted returns (=  $R_G$ )

- previous example from mwrr

$$[(1.15)(1.066)]^{1/2} - 1 = 10.75498\%$$

- large funds, HPR = 1 day  $\rightarrow (1 + HPR_1)(1 + HPR_2) + \dots + (1 + HPR_{365}) - 1$   
for liquid underlyings with market prices

c

• Annualized Return/ - all rates/returns are quoted annually

$$R_{\text{annual}} = (1 + R_{\text{period}})^{365/\text{period}} - 1 \rightarrow \text{daily } (1 + R)^{365} - 1 \quad 7\text{d/wk. } 365\text{d/yr.}$$

period = 43 days

$$R_{\text{annual}} = (1 + R_{43})^{365/43} - 1$$

$$\text{weekly } (1 + R)^{52} - 1 \quad 52\text{wk./yr.}$$

$$\text{monthly } (1 + R)^{12} - 1 \quad 12\text{mos./yr.}$$

$$\text{quarterly } (1 + R)^4 - 1$$

period = 540 days

$$R_{\text{annual}} = (1 + R_{540})^{365/540} - 1$$

250 trading days

5d/wk., 20d/m., 250d/yr.

- annualizing returns can be misleading - assumes that returns can be repeated

• Continuously Compounded Returns

d

$$\left. \begin{array}{l} \ln \left( \frac{P_t}{P_0} \right) \quad \text{e.g./ } r_c = \ln \left( \frac{105}{100} \right) = 4.879 \\ \text{or } \ln(1 + r) \quad r_c = \ln(1.05) \end{array} \right\} \begin{array}{l} e^{r_c} - 1 = r \\ e^{.04879} - 1 = 5\% \end{array}$$

• **Gross and Net Return/**

**gross** - return before deductions for mgmt. exp., custodial fees, taxes, etc. but after trading expenses (what the fund earns)

- appropriate measure for evaluating and comparing the investment skill of managers

**net** - what the investor earns

**Pre-tax and After-tax Nominal Return/**

- default is to report/state pre-tax return

- each investor's marginal tax rate may differ

- components of return may be reported e.g. 11% / 2% div.  
/ 1% int.  
/ 4% realized gains

**Real Returns/**

e  $1 + \text{real return} = \frac{1 + \text{nominal return}}{1 + \text{inflation rate}} \quad \text{---} \quad \frac{1 + \text{nominal return}}{1 + \text{risk-free rate}} = 1 + \text{risk premium}$

(typically p-o-p CPI)

**Real Returns/**

After-tax real return → investor measure of growth in purchasing power of portfolio

**Leveraged Returns/**

- leverage can be obtained through margin loans, derivatives, or collateralized loans (repos)

- if  $R_p > r_d$ , leverage enhances return

Value borrowed

$$R_L = \frac{R_P \times (V_E + V_B) - (V_B \times r_d)}{V_E} = R_P + \frac{V_B}{V_E} (R_P - r_d)$$

↑ leveraged return      ↓ portfolio equity

$$\left[ \frac{R_P V_E}{V_E} + \frac{R_P V_B}{V_E} - \frac{V_B r_d}{V_E} = R_P + \frac{V_B}{V_E} (R_P - r_d) \right]$$

e e.g./ 10m equity portfolio

$R_p = 8\%$

30% debt financed,  $r_d = 5\%$

$$R_L = 8\% + \frac{3M}{7M} (8\% - 5\%) = 9.2857\%$$

## **The Time Value of Money in Finance**

- a. calculate and interpret the present value (PV) of fixed-income and equity instruments based on expected future cash flows
- b. calculate and interpret the implied return of fixed-income instruments and required return and implied growth of equity instruments given the present value (PV) and cash flows
- c. 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

## The Time Value of Money in Finance

• 3 rules of money

- larger CFs are worth more
- less risky CFs are worth more (lower discount rate)
- CFs sooner are worth more (time value of money)



- calculate PV → value today - requires discounting at a rate that depends on the timing and type of CF

recall:  $r_f + \pi^e$  - gov't. bonds

- + default } corporate/private debt
- + liquidity } corporate/private debt
- + maturity - longer-term debt
- + equity - equity over debt

$$FV = PV(1 + r)^t \qquad \text{or} \qquad FV = PVe^{rT}$$

$$PV = \frac{FV}{(1 + r)^t} = FV(1 + r)^{-t} \qquad PV = FVe^{-rT}$$

single cash flow.

**Fixed Income/ debt instruments (bonds, loans, mortgages, etc.)**

- ZCB - zero coupon bonds (i.e. T-Bills → up to 1 yr. mat.)
  - sold at a discount, mature at par - single CF at maturity
- Coupon bonds (Notes, Bonds)
  - investor receives a number of interest payments over time and par at maturity
- Fully amortizing bonds (mortgage, auto loan)
  - investor receives level payments of both interest and principal

- ZCB/zero-coupon bond:**  $PV = \frac{FV}{(1+r)^T}$   $r =$  discount rate, IRR, or YTM
- e.g./
- 20-yr. ZCB, YTM = 6.7%  $\rightarrow PV = \frac{100}{(1.067)^{20}} = 27.33453$   
(TVM keys FV = 100 I/Y = 6.7 PMT = 0 N = 20 CPT PV)
  - price in 3 years if YTM is unchanged?
    - a)  $FV_3 = PV(1+r)^3 = 27.33453(1.067)^3 = 33.20510591$
    - or b)  $PV_3 = \frac{FV}{(1+r)^{17}} = \frac{100}{(1.067)^{17}} = 33.20510591$
  - $PV = 22.68224 \rightarrow YTM = ?$   
 $22.68224 = \frac{100}{(1+r)^{20}} \rightarrow PV = -22.68224$   
 $(100/22.68224)^{1/20} - 1 = r$   
 FV = 100 N = 20 PMT = 0 CPT I/Y = 7.6999 ~ 7.7%
  - $r = -.05$ , 10 yr. ZCB  
 $PV = \frac{100}{(.9995)^{10}} = 100.50137$  FV = 100 N = 10 I/Y = -.05 PMT = 0 CPT PV
  - 6 yrs. later  $P_0 = 95.72$ , YTM = ?  $95.72 = \frac{100}{(1+r)^4} \rightarrow (100/95.72)^{1/4} - 1 = 1.09957\%$   
 $PV = -95.72$  FV = 100 N = 4 PMT = 0 CPT I/Y

- **Coupon bond**  

$$PV = \frac{PMT_1}{(1+r)} + \frac{PMT_2}{(1+r)^2} + \dots + \frac{FV + PMT_N}{(1+r)^N}$$

FV = 100  
N = yrs.  $\times$  period  
PMT = coupon/period  
I/Y = YTM/period
- 7 yr., 2% annual bond issued at YTM = 2%  
 par bond FV = 100 PMT = 2 I/Y = 2 N = 7 CPT PV
- one year later, P = 93.091, YTM = ?  
 FV = 100 N = 6 PMT = 2 PV = -93.091 CPT I/Y = 3.287566
- 20 yr., 6.7% semi issued at YTM = 6.70% par bond
  - 7.7%? FV = 100 N = 20  $\times$  2 = 40 PMT =  $\frac{6.70}{2} = 3.35$  I/Y =  $\frac{7.70}{2} = 3.85$   
 CPT PV = 89.8788
- 20 yr. ZCB assuming YTM = 6.7% semi  
 FV = 100 N = 40 I/Y = 3.35 PMT = 0 CPT PV = 26.7658  
 vs. 27.33 annually

• Perpetuity (some bonds, preferred shares)

$$PV = \frac{PMT}{r} \rightarrow YTM$$

e.g./ 3.3% qtlly. coupon, P = 97.03

$$97.03 = \frac{.825}{r_4} \rightarrow r_4 = .8502\% \\ r = 3.401\%$$

• Annuity (mortgage, car loans)

$$A = \frac{r(PV)}{1 - (1+r)^{-t}} \quad \text{or/ TVM keys} \rightarrow \text{CPT PMT}$$

e.g./ 800k, 30yr. - FRM @ 5.25%

$$N = 30 \times 12 = 360$$

FV = 0 (fully amortizing)

$$PV = -800,000$$

$$I/Y = 5.25/12 = .4375 \quad \text{CPT PMT}$$

4,417.63

Amortization Table:

	PMT	i	P	B
Month 1	4417.63	3500	917.63	799,082.37
2	4417.63	3495.99	921.64	798,160.73

$\downarrow$   
 $B \times r/12$

$\backslash$   
 $PMT - i$

$\backslash$   
 $B_0 - P = B_1$

• Equity - pref. shares, common shares

- constant dividend - perpetuity
- growing dividend (constant rate) - growing perpetuity
- growing dividend (non-constant rate)

Constant dividend/ - many REITs

- perpetuity

$$PV = D/r$$

$$D = 1.50 \quad r = 15\%$$

$$PV = 1.50 / .15 = 10$$

Constant growth dividend - commercial real estate (to calculate

property values)

$$PV = \frac{D_0(1+g)}{r-g}$$

$$\text{for CRE: Prop. Value} = \frac{NOI_1}{r-g}$$

$\rightarrow$  cap rate

Variable growth dividend - growth moving to value

- 2 stage model

$$PV = \sum_{i=1}^n \frac{D_0(1+g_s)}{(1+r)^i} + \frac{D_0(1+g_s)^n(1+g_L)}{(1+r)^n} \cdot \frac{1}{r-g}$$

explicit discount period

$\nearrow$  terminal value = perpetuity

e.g./  $D_0 = 1.50$   $g = 6\%$  - growing perpetuity (typical of value stocks)  
 $r = 15\%$

$$PV = \frac{1.50(1.06)}{.15 - .06} = 1.59 / .09 = 17.67$$

→ now assume  $g_s = 6\%$  for 3 yrs.,  $g_L = 2\%$  thereafter

$$PV = \frac{1.50(1.06)}{1.15} + \frac{1.50(1.06)^2}{(1.15)^2} + \frac{1.50(1.06)^3}{(1.015)^3} + \frac{1.50(1.06)^3(1.02)}{.15 - .02}$$

$$= 1.3826 + 1.2744 + 1.1747 + \frac{14.01734}{(1.15)^3}$$

$$= 13.04833$$

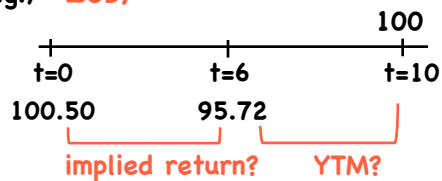
recall:  $PV = \frac{FV}{(1+r)^T} \rightarrow PV(1+r)^T = FV$

$$(1+r)^T = FV/PV$$

$$1+r = (FV/PV)^{1/T}$$

$$r = (FV/PV)^{1/T} - 1$$

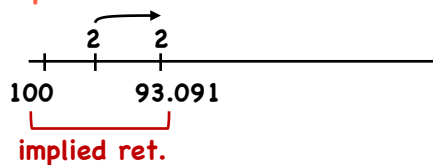
e.g./ ZCB/



$$r = \left( \frac{95.72}{100} \right)^{1/6} - 1 = -.8088\%$$

$$r = \left( \frac{100}{95.72} \right)^{1/4} - 1 = 1.10\%$$

Coupon bond/



$$FV = ? = 2(1.02) + 2 + 93.091$$

$$= 97.131$$

$$r = \left( \frac{97.131}{100} \right)^{1/2} - 1 = -1.445\%$$

Equity/  $PV = \frac{D_0(1+g)}{r-g} \rightarrow PV(r-g) = D_0(1+g)$

e.g./  $P_0 = 63 \quad D_1 = 1.76 \quad g = 4\%$

$r = \frac{1.76}{63} + 4\% = 6.79\%$

- what if  $r = 7\% \rightarrow g?$

$g = 7\% - \frac{1.76}{63} = 4.21\%$

$r - g = \frac{D_0(1+g)}{PV}$

$r = \frac{D_0(1+g)}{PV} + g$  → GGM  
div. yield      growth

$g = r - \frac{D_0(1+g)}{PV}$

•  $PV = \frac{D_0(1+g)}{r-g} \rightarrow$  divide both sides by E

$\frac{PV}{E} = \frac{D_0/E(1+g)}{r-g}$   
PE ratio

→ dividend payout ratio - DPR

- given a  $P/E$  and DPR - we can solve for r or g given the other

- compare the implied r or g to the required r or expected/estimated g

forward P/E →  $\frac{PV}{E_{t+1}} = \frac{D_{t+1}/E_{t+1}}{r-g}$

e.g./ 1/ forward PE = 28      E(DPR) = 70%

$g = 4\% \quad r = ?$

$28 = \frac{70\%}{r - .04} \rightarrow 28r - 1.12 = .7$   
 $28r = 1.82$

$r = 1.82/28 = 6.5\%$

- if the forward div. is expected to ↑ or g is expected to ↑, stock/index will trade at a higher forward multiple

- a higher required return leads to a lower multiple

2/ forward PE = 19      E(DPR) = 60%

$r = 8\% \quad g = ?$

$19 = \frac{.60}{.08 - g} \rightarrow 1.52 - 19g = .6$   
 $.92 = 19g$

$g = .92/19 = 4.84\%$

3/ forward PE = 28      E(DPR) = 70%  
 $r = 9\% \quad g = 4.5\%$  buy/sell stock?

$PE = \frac{.7}{.09 - .045} = 15.5 \times$

- should be trading at 15.5× forward earnings, not 28 ∴ sell

• Cash flow additivity → principle of no arbitrage - 2 economically equivalent strategies should have the same price

$r = 6\%$   
which one?

A	45	45	45
B	60	40	32.5
A - B	-15	5	12.50

- calculate PV of both, select the higher one  
or/ take the difference in CFs (A - B)  
if  $PV > 0$ , choose A, else B

$PV = 0$  → both equivalent

Proof/

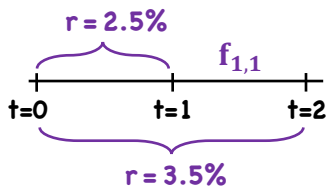
$r = 6\%$

A	100
B	120
	-20
	$-20/1.06 = -18.8679$

- B is clearly superior

$PV(A) = 94.339$   
 $PV(B) = 113.2075$   
 $A - B = -18.8679$

$-20/1.06 = -18.8679$



- invest for 2 yrs.
  - buy a 2 yr. ZCB
  - buy a 1 yr. ZCB, buy another in one year
- |  
spot
- |  
 $f_{1,1}$   
(implied forward rate)

$(1.035)^2 = (1.025)(1 + f_{1,1})$   
 $f_{1,1} = \frac{(1.035)^2}{(1.025)} - 1 = 4.5097\%$

- if  $f_{1,1} = 5\%$  → lock-in rate with FRA →  $(1.025)(1.05) = 1.07625$

- would provide an arbitrage profit of .5025 per 100 of par

vs.  
 $(1.035)^2 = 1.071225$   
 $.005025$

e.g./

	1 yr.	2 yr.	$\Delta f_{1,1} = ?$	
May 31	98.028	95.109	$(100/98.028) - 1 = 2.011\%$	$f_{1,1} = \frac{(1.02539)^2}{(1.02011)} - 1 = 3.069\%$
June 15	97.402	93.937	$(100/95.109)^{1/2} - 1 = 2.539\%$	

e.g./	1yr.	2yr.	$\Delta f_{1,1} = ?$
May 31	98.028	95.109	$f_{1,1} = 3.069\%$
June 15	97.402	93.937	$\rightarrow (100/97.402) - 1 = 2.6673\%$ $(100/93.937)^{1/2} - 1 = 3.1767\%$

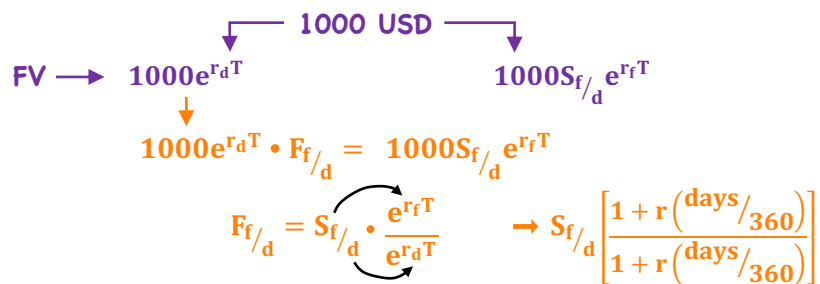
$$\frac{(1.031767)^2}{(1.026673)} - 1 = 3.6886$$

$\Delta f_{1,1} = 3.6886\% - 3.069\% = 61.964 \text{ bps}$

or/  $98.028/95.109 = 3.069\%$   
 $97.402/93.937 = 3.688\%$

$3.688\% - 3.069\%$

forex - forward  
exchange rate  
 $F_{f/d}$



Forex/ e.g./  $S_{JPY/USD} = 134.40$   $r_{USD} = 2\%$   $r_{JPY} = .05\%$  6 mos.  $F_{JPY/USD} = ?$

$F_{JPY/USD} = S_{JPY/USD} \cdot \frac{e^{r_{JPY} T}}{e^{r_{USD} T}} = 134.40 \left( \frac{e^{0.0005(5)}}{e^{0.02(5)}} \right) = 133.0959$

e.g./  $r_f$   $r_d$

May 31 2.012% 1.291%  $\rightarrow F_{f/d} = 1.2602 \cdot \frac{e^{0.0212}}{e^{0.01291}} = 1.269319$

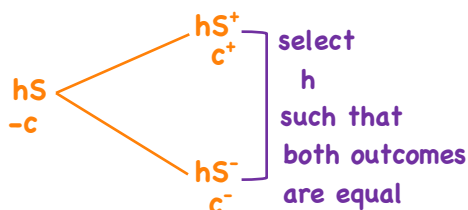
June 15 2.667% 1.562%  $\rightarrow F_{f/d} = 1.2602 \cdot \frac{e^{0.02667}}{e^{0.01562}} = 1.274202$

$S_{f/d} = 1.2602$  1 yr.  $\rightarrow \Delta F_{f/d} = ?$

$\Delta F_{f/d} = \uparrow .004883$

or 48.83 pips

Option pricing/



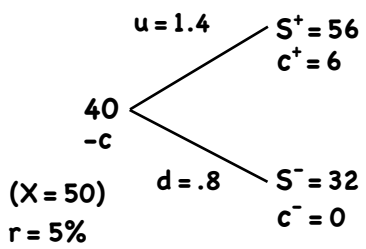
$hS^+ - c^+ = hS^- - c^-$

$hS^+ - hS^- = c^+ - c^-$

$h(S^+ - S^-) = c^+ - c^-$

$h = \frac{c^+ - c^-}{S^+ - S^-}$

Option pricing/



$$h = \left| \frac{6 - 0}{56 - 32} \right| = 6/24 = .25$$

$$hS^+ - c^+ = .25(56) - 6 = 8$$

$$hS^- - c^- = .25(32) - 0 = 8$$

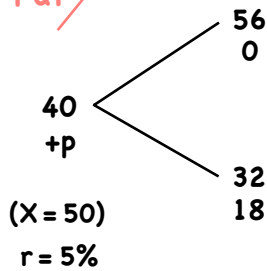
identical  
 $\therefore 0.25S - c$  is  
 a risk-free portfolio

$$hS_0 - c_0 = \frac{hS^+ - c^+}{1 + r}$$

$$.25(40) - c_0 = \frac{.25(56) - 6}{1.05}$$

$$c_0 = 10 - 8/1.05 = 2.38095$$

Put



$$h = \left| \frac{0 - 18}{56 - 32} \right| = .75$$

$$hS_0 + p_0 = \frac{hS^- + p^-}{1 + r}$$

$$.75(40) + p_0 = \frac{.75(32) + 18}{1.05}$$

$$p_0 = 42/1.05 - 30$$

$$p_0 = 10$$

## **Statistical Measures of Asset Returns**

- a. calculate, interpret, and evaluate measures of central tendency and location to address an investment problem
- b. calculate, interpret, and evaluate measures of dispersion to address an investment problem
- c. interpret and evaluate measures of skewness and kurtosis to address an investment problem
- d. interpret correlation between two variables to address an investment problem

## Statistical Measures of Asset Returns

- Measures of central tendency/
  - where the data are centered i.e. the expected value
- Arithmetic mean  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ 
  - describes a representative possible outcome
  - sensitive to outliers (extreme values)
- Median - middle value  $\frac{(n+1)}{2}$  odd # of obs.
  - not affected by outliers  $\frac{(\frac{n}{2}) + (\frac{n+2}{2})}{2}$  even # of obs. - average the middle 2 obs.
- mode - most frequently occurring value in a dataset
  - a dataset can have more than one mode, or no mode
  - single mode → unimodal all obs. are different
  - two modes → bimodal

Return Bin (%)	↓ Absolute Frequency	↓ Relative Frequency (%)	↓ Cumulative Absolute Frequency	↓ Cumulative Relative Frequency (%)
-5.0 to -4.0	1	0.08	1	0.08
-4.0 to -3.0	7	0.56	8	0.64
-3.0 to -2.0	23	1.83	31	2.46
-2.0 to -1.0	77	6.12	108	8.59
-1.0 to 0.0	470	37.36	578	45.95
0.0 to 1.0	555	44.12	1,133	90.06
1.0 to 2.0	110	8.74	1,243	98.81
2.0 to 3.0	13	1.03	1,256	99.84
3.0 to 4.0	1	0.08	1,257	99.92
4.0 to 5.0	1	0.08	1,258	100.00
	count	%	Σcount	Σ%

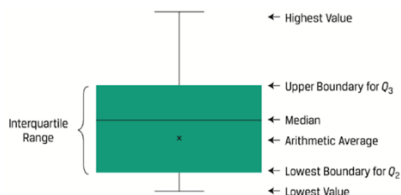
- Outliers/**
- do nothing - if values are legitimate and correct
  - eliminates judgment
  - Delete → trimmed mean
  - Replace → winsorized mean
  - compare mean with and without outliers

## Measures of location - quantiles

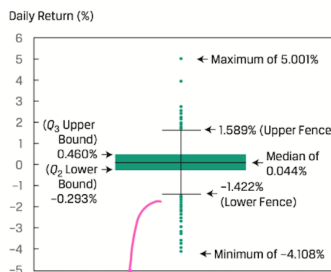
## IQR - interquartile range

- quartiles 25, 50, 75
- quintiles 20, 40, 60, 80
- deciles 10, 20 ... 90
- percentiles 1, 2 ... 99

### • Box and whisker plot



Box & whisker plot



upper = (1.5 x IQR)  
+ upper bound

lower fence = lower bound - (1.5 x IQR)

- uses/
- rank performance of portfolios and investment managers in terms of percentile/quartile in which they fall

- investment research → bottom return decile → short } long/short HF
- top return decile → long }

- Dispersion - variability around the central tendency  
- a measure of risk or uncertainty

- range: max. value - min. value  
- no information about the shape of the distribution  
- sensitive to outliers

- mean absolute deviation  $MAD = \frac{\sum |X_i - \bar{X}|}{n}$

- sample variance

in units squared  
- can be added/subtracted

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

degrees of freedom  
(n - 1 independent obs.)

$$\text{pop: } \sigma^2 = \frac{\sum_{i=1}^n (X - \mu)^2}{n}$$

- sample standard deviation

- in the same units as the data itself

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}$$

$$\text{pop } \sigma = \sqrt{\sigma^2}$$

• Downside deviation

e.g. target semideviation:

$$S_{\text{target}} = \sqrt{\frac{\sum_{i=1}^n (X_i - B)^2}{n - 1}}$$

full sample n

• as B ↑ S<sub>target</sub> ↑

∀ X<sub>i</sub> ≤ B  
|  
some minimum level

• coefficient of variation

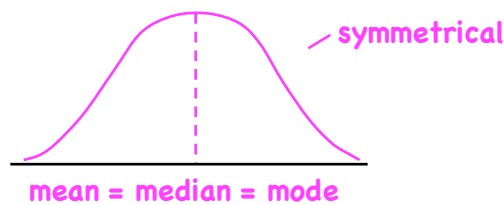
$$CV = \frac{S}{\bar{X}}$$

- a measure of relative dispersion

e.g. - for returns, CV measures the risk per unit of return

- lower = better (less uncertainty, tighter distribution)

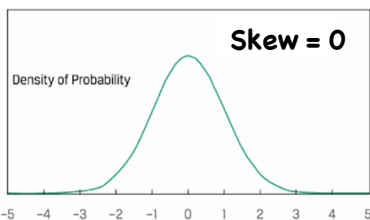
• Normal distribution - most common



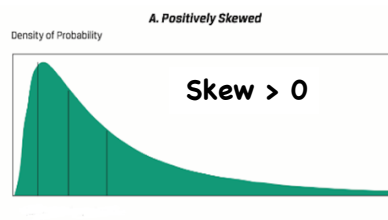
• completely described by 2 parameters:  $\mu, \sigma^2$

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

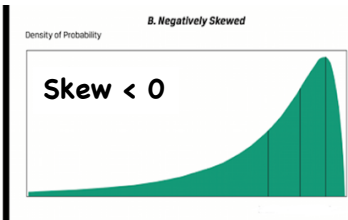
Skew



mean = median = mode



mean > median > mode

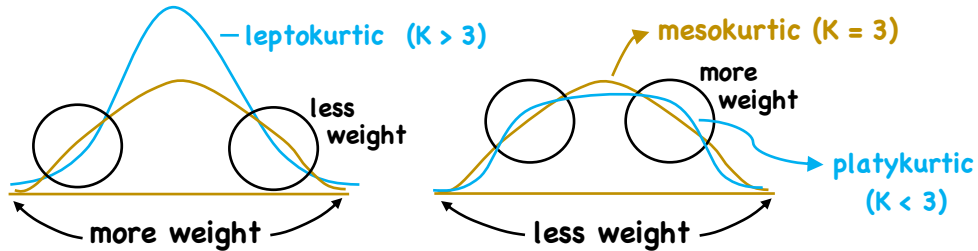


mean < median < mode

$$\text{Skew} \approx \frac{1}{n} \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{S^3} \quad \text{for } n > 100$$

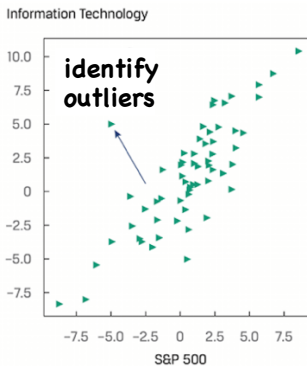
$$\text{Kurtosis } K_E = \left[ \frac{1}{n} \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{S^4} \right] - 3$$

- measures the combined weight of the tails relative to the rest of the distribution



- Lepto →  $K_e > 0$  - ow head, uw shoulders, ow tails
- Meso →  $K_e = 0$  - normal dist.
- Platy →  $K_e < 0$  - uw head, ow shoulders, uw tails

### Scatter Plot



- used to visualize the joint variation in 2 numerical values

- may be no relationship, a linear or non-linear relationship

- scatter plot matrix

- assess for pairwise association among many variables

**Covariance** → the joint variability of 2 random variables  
→ expressed in the same units as the variables

$$S_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$S_{XY} > 0$  when they covary together  
 $(X_i - \bar{X}) > 0$  when  $(Y_i - \bar{Y}) > 0$   
and  $(X_i - \bar{X}) < 0$  when  $(Y_i - \bar{Y}) < 0$

**Correlation** → measures the **linear** association between 2 variables

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

**Properties:**

• **Cov(XY)**  
determines the  
sign of  $r_{XY}$

1/  $-1 \leq r \leq 1$

2/  $r = 0$  implies no linear relationship → maximum diversification

3/  $r = 1$  → perfect positive correlation → perfect replication

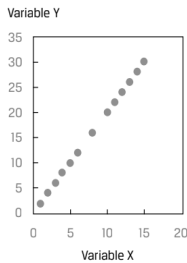
4/  $r = -1$  → perfect negative correlation → perfect hedge

**Limitations/**

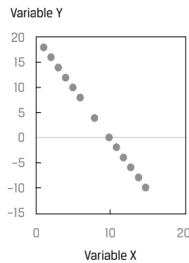
- linear association only
- unreliable when outliers are present
- correlation does not imply causation
  - spurious correlation → chance relationship
    - a third variable



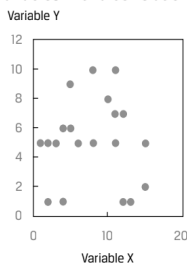
**A. Variables With a Correlation of +1**



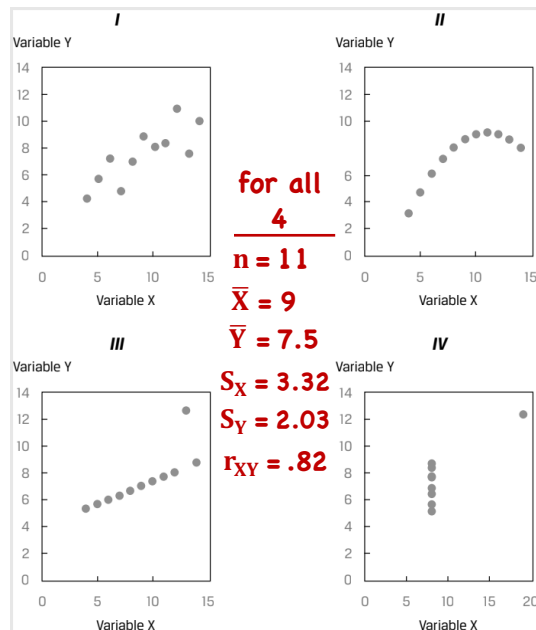
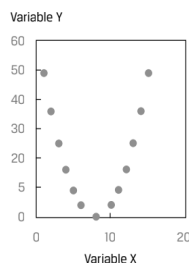
**B. Variables With a Correlation of -1**



**C. Variables With a Correlation of 0**



**D. Variables With a Strong Nonlinear Association**



∴ visual inspection important

## **Probability Trees and Conditional Expectations**

- a. calculate expected values, variances, and standard deviations and demonstrate their application to investment problems
- b. formulate an investment problem as a probability tree and explain the use of conditional expectations in investment application
- c. calculate and interpret an updated probability in an investment setting using Bayes' formula

## Probability Trees and Conditional Expectations

- Key point - investment decisions are made under uncertainty

- the expected value of a random variable  $\rightarrow E(X) \rightarrow$  is a probability-weighted average of the possible outcomes

$E(X)$  — forecast of future value  
           — estimate of the 'true' population mean based on a sample

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

↓  
one outcome only

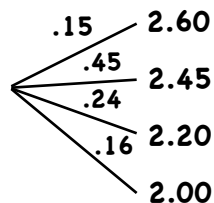
- Variance of a random variable

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

$$\sigma(X) = \sqrt{\sigma^2(X)}$$

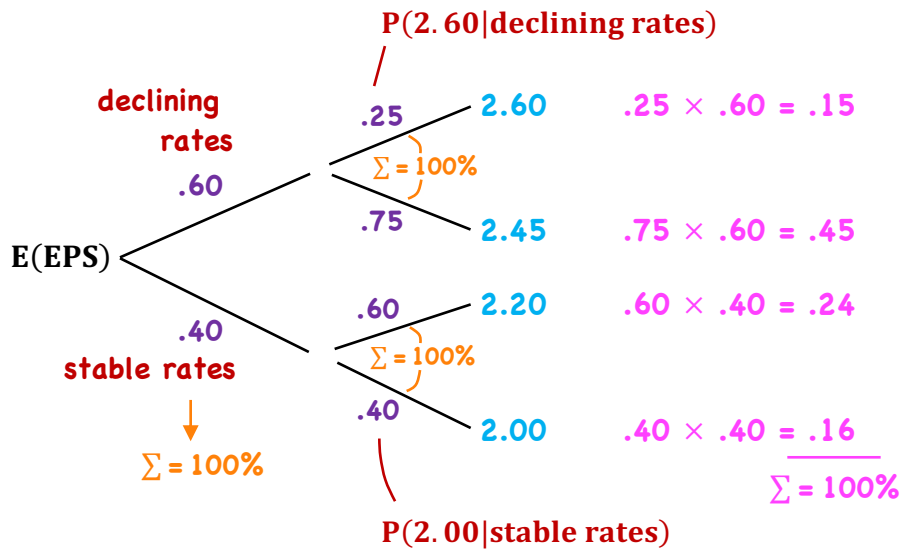
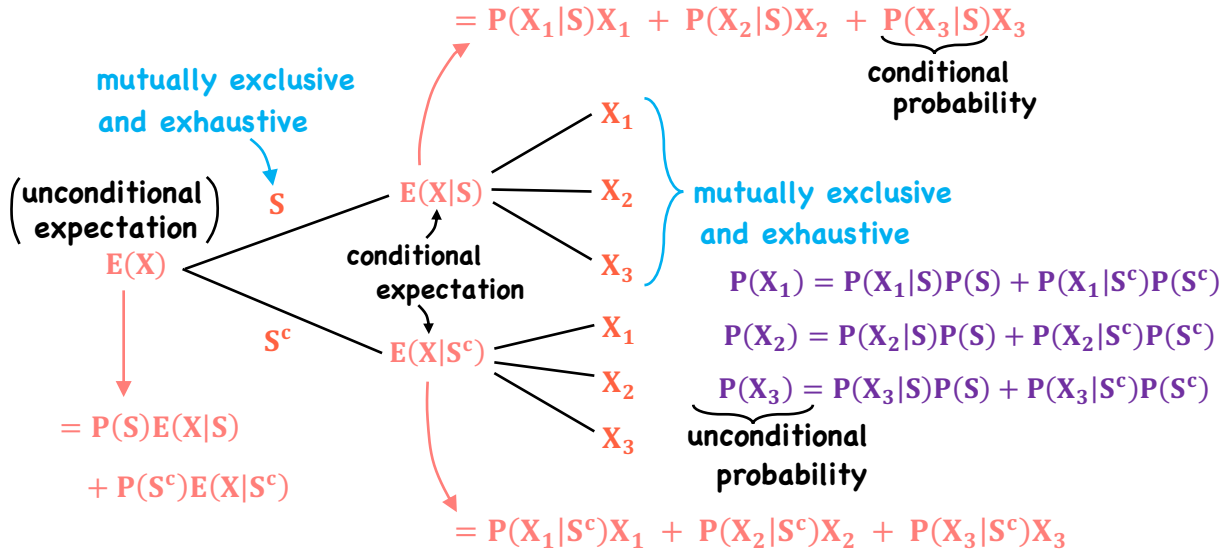
e.g./	P	EPS	(X)
	.15	2.60	
	.45	2.45	
	.24	2.20	
	.16	2.00	

$$E(X) = .15(2.60) + .45(2.45) + .24(2.20) + .16(2.00) = 2.3405$$



$$\sigma^2(X) = .15(2.60 - 2.3405)^2 + .45(2.45 - 2.3405)^2 + .24(2.20 - 2.3405)^2 + .16(2.00 - 2.3405)^2 = .038785$$

$$\sigma(X) = \sqrt{.038785} = .196939$$



$$E(EPS) = P(X_1|S)P(S)X_1 + P(X_2|S)P(S)X_2 + P(X_3|S^c)P(S^c)X_3 + P(X_4|S^c)P(S^c)X_4$$

$$(.25)(.60)2.60 + (.75)(.60)2.45 + (.60)(.40)2.20 + (.40)(.40)(2.00)$$

- **Bayes' Formula:** a method for updating prior probabilities based on new information

**Recall: Total Probability Rule**

$$P(E) = P(E|S_1)P(S_1) + P(E|S_2)P(S_2) + \dots + P(E|S_n)P(S_n)$$

Q: given that we observe E, what is  $P(S_n)$ ? →  $P(S_n|E)$

$$P(S_n|E) = \frac{P(E|S_n)P(S)}{P(E)} = \frac{P(E|S_n)}{P(E)} \cdot P(S)$$

e.g. /

prior ↓

Exp. .75

no Exp. .25

Exp. .20

no Exp. .80

Exp. .05

no Exp. .95

$P(E) =$

$.45(.75) (.3375)$

$+$

$.30(.20) (.06)$

$+$

$.25(.05) (.0125)$

$41\%$

What is  $P(\text{exceeds}|\text{expands})$

$= \frac{(.75)(.45)}{.41} = \frac{.75}{.41} \cdot .45 = .8231$

$= \frac{P(\text{Exp.}|\text{Exc.})P(\text{Exc.})}{P(\text{Exp.})} = \frac{P(\text{Exp.}|\text{Exc.})}{P(\text{Exp.})} \cdot P(\text{Exc.})$

## **Portfolio Mathematics**

- a. calculate and interpret the expected value, variances, standard deviation, covariances, and correlations of portfolio returns
- b. calculate and interpret the covariance and correlation of portfolio returns using the joint probability function for returns
- c. define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy's safety-first criterion

## Portfolio Mathematics

**calculate and interpret the expected value,  
variance, standard deviation, covariances,  
and correlations of portfolio returns**

⇒ **Portfolio Returns (  $E(R_P)$ ,  $\sigma_{R_P}$ ,  $Cov_{ij}$ ,  $\rho_{ij}$  )**

$$1/ E(R_P) = E(W_1R_1 + W_2R_2 + \dots + W_nR_n)$$

↓  
also a random variable

$$E(R_1) = \underbrace{P(R_{11})}_{\text{probability}}R_{11} + P(R_{12})R_{12} + \dots + P(R_{1n})R_{1n}$$

↗ possible value of  $R_1$

e.g./	<u>W</u>	<u>E(R<sub>i</sub>)</u>
SnP500	.50	13%
Corp. bonds	.25	6%
MSCI EAFE	.25	15%

$$E(R_P) = .5(13\%) + .25(6\%) + .25(15\%) = 11.75\%$$

↑  
measure of expected reward

$$2/ \sigma^2(R_P) = \sum_{i=1}^n \sum_{j=1}^n W_i W_j \underbrace{Cov(R_i R_j)}$$

↓  
to calculate  
portfolio variance, need:

- 1/ all  $E(R_i)$
- 2/ all  $Cov(R_i, R_j)$

**(Exhibit #11)**

$$Cov(R_i R_j) = \frac{\sum_{t=1}^n (R_{it} - \bar{R}_i)(R_{jt} - \bar{R}_j)}{n - 1}$$

↗  $E(R_i) = \sum_{t=1}^n P(R_i)R_i$

• assume 3 assets →  $R_1, R_2, R_3$

$$\begin{aligned} \sigma^2(R_P) = & W_1^2 \sigma^2(R_1) + W_2^2 \sigma^2(R_2) + W_3^2 \sigma^2(R_3) \\ & + 2W_1 W_2 Cov(R_1 R_2) + 2W_1 W_3 Cov(R_1 R_3) \\ & + 2W_2 W_3 Cov(R_2 R_3) \end{aligned}$$

•  $\sigma^2(R_p) = f(\text{variances, covariances})$

always  $> 0$       can be  $< 0$  or  $> 0$

- major point - by selecting assets with zero or negative covariance, portfolio risk is lowered

- for  $n$  securities (or asset classes)  $\rightarrow n$  variances

$n^2 - n$  covariances

$(n^2 - n)/2$  distinct covariances  $20/2 = 10$  unique Covars.

$n = 5$

5 vars.

$25 - 5 = 20$  Covars.

3/ Correlation  $\rightarrow \rho_{ij} = \frac{\text{Cov}(R_i R_j)}{\sigma_{R_i} \cdot \sigma_{R_j}}$

$\rightarrow \text{Cov}(R_i R_j) = \rho_{ij} \cdot \sigma_{R_i} \cdot \sigma_{R_j}$

Exhibit 12 Covariance Matrix				Exhibit 13 Correlation Matrix of Returns		
	S&P 500	US Long-Term Corporate Bonds	MSCI EAFE	S&P 500	US Long-Term Corporate Bonds	MSCI EAFE
S&P 500	400	45	189	1.00	0.25	0.45
US long-term corporate bonds	45	81	38	0.25	1.00	0.20
MSCI EAFE	189	38	441	0.45	0.20	1.00

$\frac{189}{400^{1/2} \cdot 441^{1/2}} = \frac{189}{20.21} = 0.45$  **Ex #13**

Recall:

$\text{Cov}(R_A R_B) = \frac{\sum (R_{A_i} - \bar{R}_A)(R_{B_i} - \bar{R}_B)}{n - 1}$

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

$= \frac{1}{n-1} (R_{A_1} - \bar{R}_A)(R_{B_1} - \bar{R}_B) + \frac{1}{n-1} (R_{A_2} - \bar{R}_A)(R_{B_2} - \bar{R}_B) + \dots + \frac{1}{n-1} (R_{A_n} - \bar{R}_A)(R_{B_n} - \bar{R}_B)$

weights

probabilities?

- the concept of joint probability

$\text{Cov}(R_A R_B) = \sum_{i=1}^n \sum_{j=1}^n P(R_{A_i} R_{B_j}) (R_{A_i} - \bar{R}_A)(R_{B_j} - \bar{R}_B)$

where  $i$  &  $j = 1$  to  $n$  are scenarios

probability

value of cross product

if returns are independent:

$\rightarrow P(R_A R_B) = P(R_A)P(R_B)$

$\rightarrow$  since independence is a stronger property than uncorrelatedness, this property holds for uncorrelated random variables

Exhibit #6/7

- Safety first rules focus on shortfall risk - the risk a portfolio value (or return) will fall below some minimum acceptable level over some time horizon

Let  $R_L$  = minimum acceptable level of return

$$\text{z-value} \rightarrow \text{SFRatio} = \frac{E(R_P) - R_L}{\sigma_P}$$

- objective is to maximize this ratio
- optimal portfolio minimizes  $N(-\text{SFRatio})$

e.g./  $R_L = 2\%$

Portfolio	$R_P$	$\sigma$
1	12%	15%
2	14%	16%

$$\text{SFRatio}_1 = \frac{12 - 2}{15} = 0.667 = \text{NORM.S.DIST}(-.667, 1) = 0.2525$$

$$\text{SFRatio}_2 = \frac{14 - 2}{16} = 0.75 = \text{NORM.S.DIST}(-.75, 1) = 0.227$$

Note: if  $R_L = R_f \rightarrow \text{SFRatio} = \text{Sharpe Ratio}$

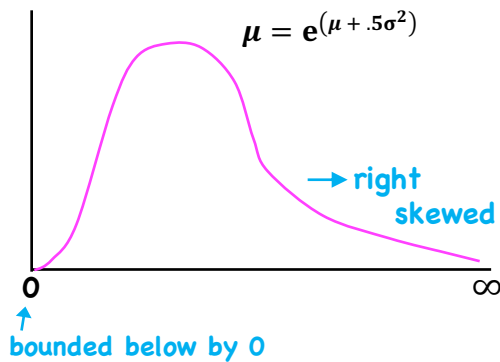
Example 3

## **Simulation Methods**

- a. 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
- b. describe Monte Carlo simulation and explain how it can be used in investment applications
- c. describe the use of bootstrap resampling in conducting a simulation based on observed data in investment applications

## Common Probability Distributions

Page 1



- commonly used to model the probability distribution of asset prices

- a variable  $Y$  follows a lognormal distribution if  $\ln(Y)$  is normally distributed

- completely described by 2 parameters  $\rightarrow$  the  $\mu$  and  $\sigma^2$  of its associated normal distribution

$$S_T/S_0 = 1 + R_H \quad \text{where } R_H = \text{holding period return}$$

$$\text{i.e./ } \frac{S_T - S_0}{S_0} = R_H$$

$$S_T/S_0 - S_0/S_0 = R_H \rightarrow S_T/S_0 - 1 = R_H \rightarrow S_T/S_0 = 1 + R_H$$

Page 2

e.g.  $S_T = 34.50 \quad S_0 = 30$

$$S_T/S_0 = 34.50/30 = 1.15 = 1 + R_H \quad \therefore R_H = 15\%$$

$$\ln(S_T/S_0) = r \quad \text{where } r = \text{continuously compounded return}$$

$$r = \ln(34.50/30) = \ln(1.15) = 0.13976 \quad \text{and } r \sim N(\mu T, \sigma^2 T)$$

$$\therefore 34.50 = 30e^{0.13976}$$

• more generally:  $S_T = S_0 e^r \quad (S_T/S_0 = e^r \text{ and } \ln(S_T/S_0) = r)$

- to assume returns are normally distributed, we assume returns are 1/ independent

2/ identically distributed ( $\mu$  and  $\sigma^2$  do not change from period to period)

- so, while  $S_T = S_0(1 + R_H)^T$

with cont. comp:  $S_T = S_0 e^{rT}$

**Volatility** → annualized sd of the continuously compounded daily returns of the underlying asset

- since  $r \sim N(\mu T, \sigma^2 T)$  ,  $sd = \sigma\sqrt{T}$

- so both the mean and variance of  $r$  scale linearly with time, but the s.d. scales linearly with the square root of time

e.g. if daily vol. = .01, annualized vol. =  $.01\sqrt{250} = 15.81\%$  example #1

### Monte Carlo Simulation/

**Step 1: Specify the quantity of interest**

e.g.  $MV_p$  in 10 years

**Step 2: Specify a time grid** →  $K$  sub-periods with  $\Delta t$  increment for the full time horizon

e.g. 20 sub-periods,  $\Delta t = 6$  months

**Step 3: Specify distributional assumptions for the key risk factors**

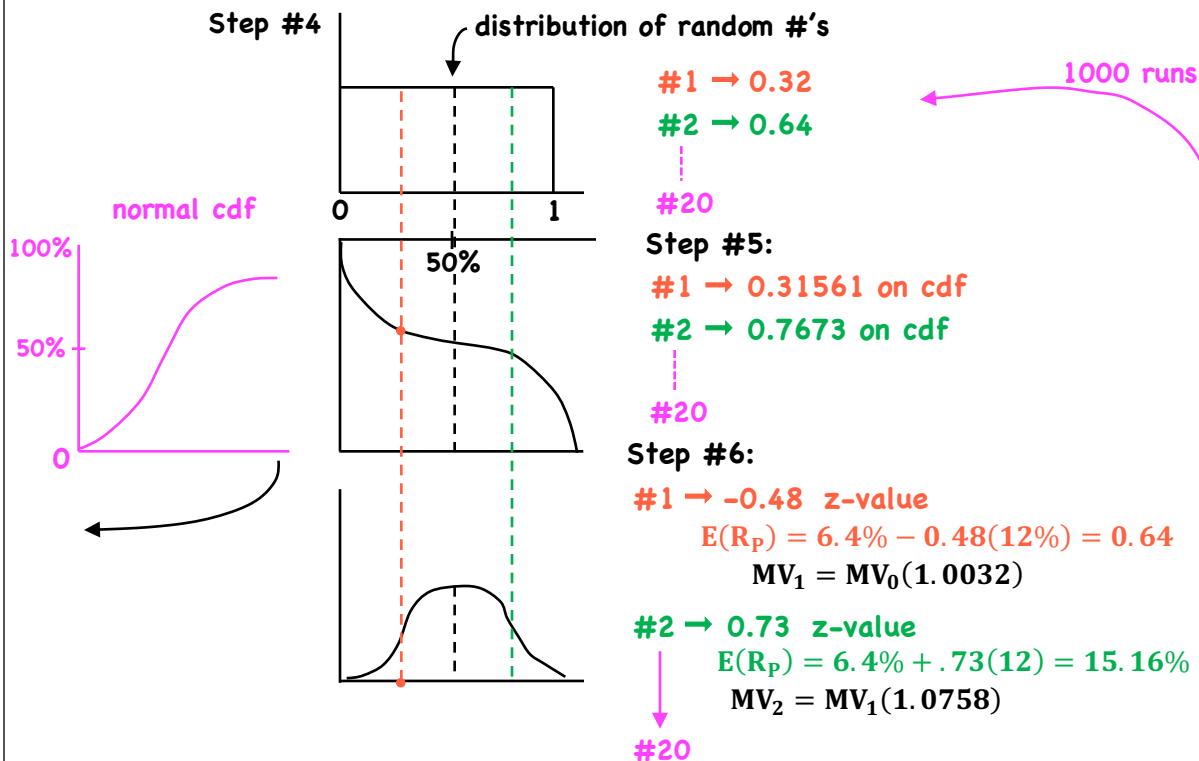
e.g.  $E(R_p) = 6.4\%$   $\sigma = 12\%$

$$MV_t = MV_{t-1}(1 + R_{p1})$$

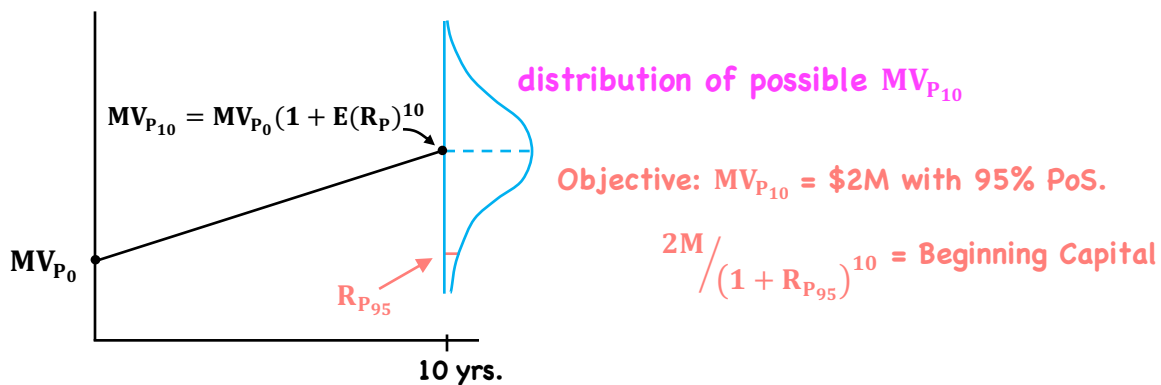
**Step 4: Draw standard normal random numbers for each key risk factor over each  $K$  sub-periods.**

- random number generator → produces a distribution of random numbers from 0 to 1, all equally likely

Monte Carlo Simulation/



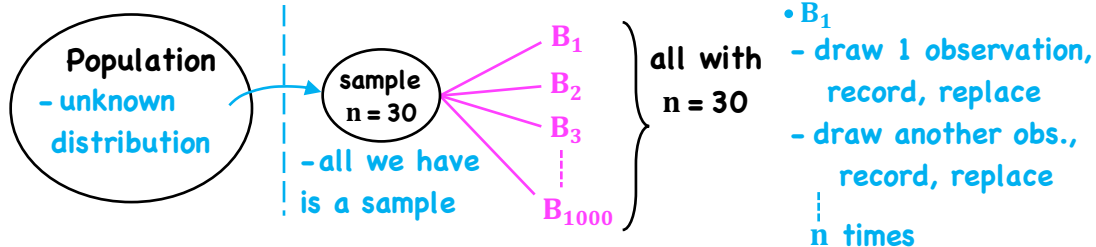
Monte Carlo Simulation/



- Monte Carlo Simulation → provides only statistical estimates, not exact results
- does not support cause and effect conclusions

**Resampling** → repeatedly draw samples from an original data sample in order to estimate population parameters

1/ **Bootstrap method**/ uses computer simulation



- rather than estimate the distribution, this method creates the distribution
- can also find SE of an estimator even when no analytical formula is available (e.g. median)

$$S_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^B (\hat{\theta}_i - \bar{\theta})^2}{B - 1}}$$

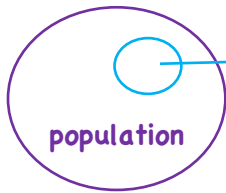
standard error

## **Estimation and Inference**

- a. compare and contrast simple random, stratified random, cluster, convenience, and judgmental sampling and their implications for sampling error in an investment problem
- b. explain the central limit theorem and its importance for the distribution and standard error of the sample mean
- c. describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

## Sampling and Estimation

Page 1

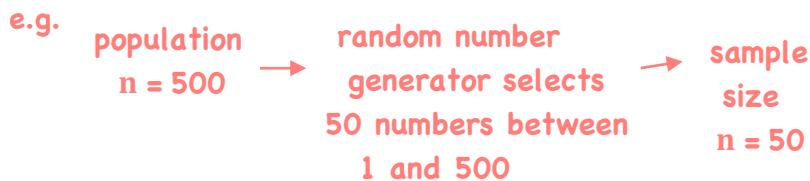


**sample** → a method of obtaining information about a population's parameters ( $\mu$  &  $\sigma$ ) through sample statistics ( $\bar{X}$  &  $S$ )

**A/ Probability sampling** → every member of a population has an equal chance of being selected

∴ samples will be more representative of the population

**1/ Simple Random Sampling** → a subset of a larger population such that each element has an equal probability of being selected



- useful when data are homogeneous

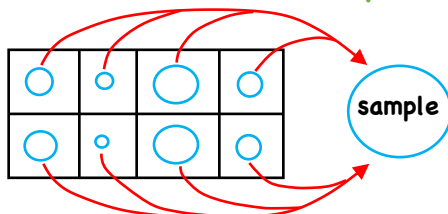
Page 2

**2/ Systematic Sampling** → when the population is too large to code

- select every  $K^{\text{th}}$  element until the desired sample size is reached

**3/ Stratified Random Sampling** - population is sub-divided into sub-populations based on one or more classifications

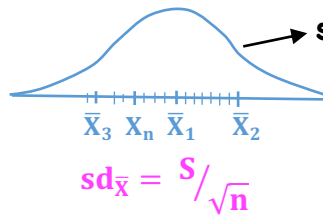
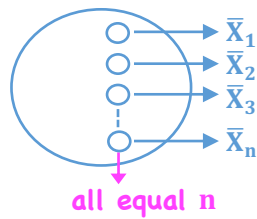
- simple random samples are then drawn from each sub-pop.
- each sample is then pooled to form the main sample



- each sub-sample is proportionate to the size of its sub-population
- guarantees that population sub-divisions are represented in the sample
- statistics will be more precise

- sample statistics are estimates of population parameters
- not exact, subject to error

• sampling error → difference between observed values of a statistic and population parameters as a result of using just a subset of the population



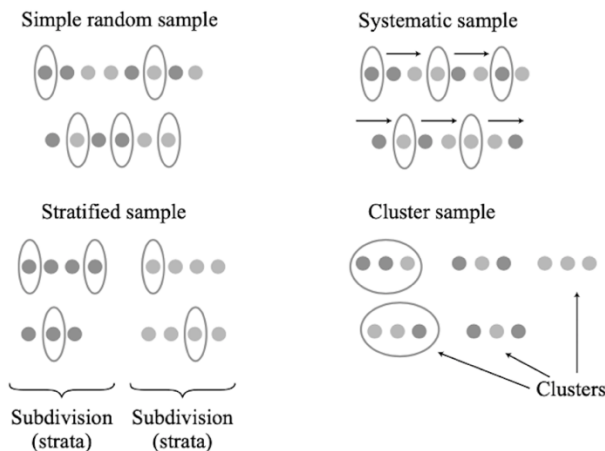
sampling distribution of the sample means

**example #1**

- 4/ Cluster sampling** - pop. is divided into clusters each of which is a mini representation of the population
- certain clusters are then selected as a whole using simple random sampling → one-stage cluster sampling

**4/ Cluster sampling** - if sub-samples are selected from each cluster → two-stage cluster sampling

- usually results in lowest precision since a cluster may not be representative of the population
- is both cost and time efficient however



**B/ Non-probability sampling** - depends on factors such as judgment or convenience (in terms of access to data)

- risk that samples may be non-representative

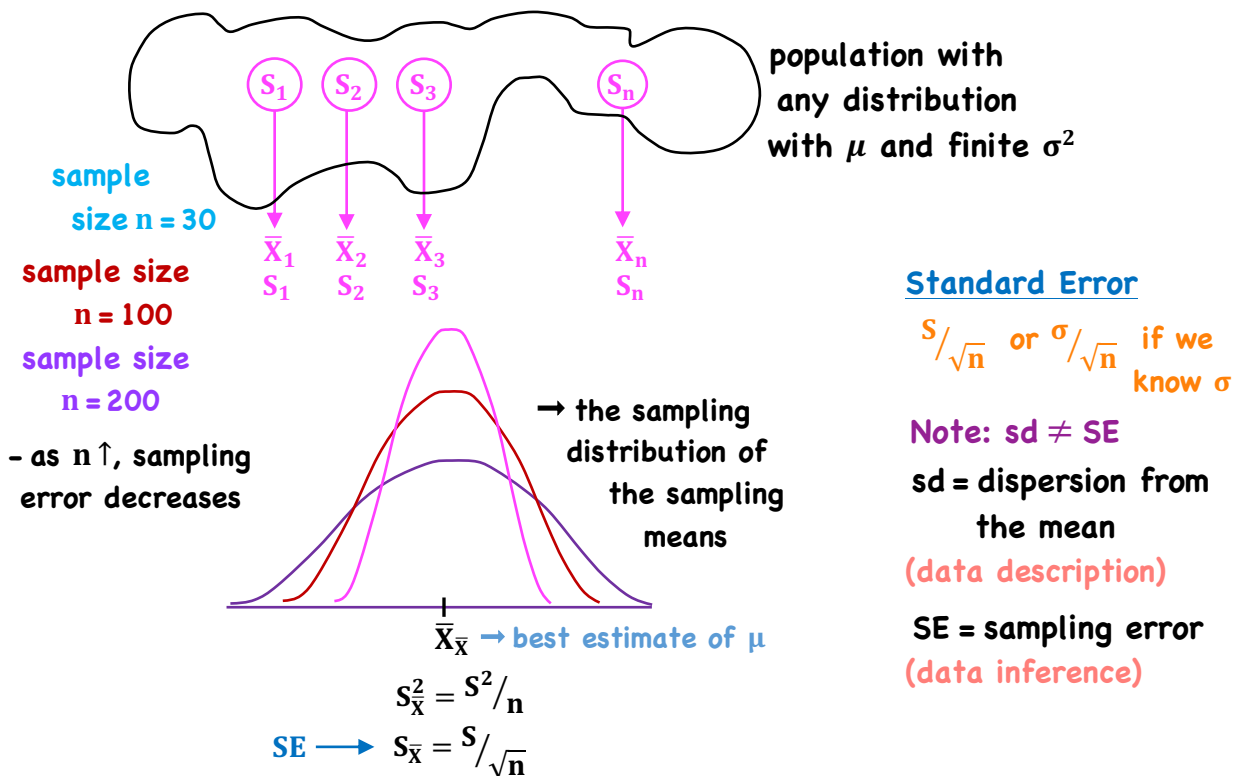
**5/ Convenience sampling** - observations are selected that are easy to obtain or are accessible

• not necessarily representative, but low cost

**6/ Judgmental sampling** → select observations based on experience and knowledge

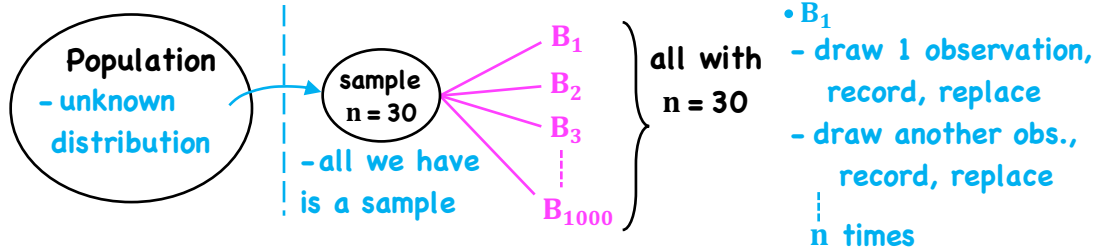
→ useful when there is a time constraint and/or the specialty of the researcher would result in better representation

example 2, 3, 4



**Resampling** → repeatedly draw samples from an original data sample in order to estimate population parameters

**1/ Bootstrap method/** uses computer simulation



- rather than estimate the distribution, this method creates the distribution
- can also find SE of an estimator even when no analytical formula is available (e.g. median)

$$S_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^B (\hat{\theta}_i - \bar{\theta})^2}{B - 1}}$$

standard error

**2/ Jackknife method/** - omit one observation from a sample, one at a time

e.g./ n = 30

J <sub>1</sub>	n = 29, omit X <sub>1</sub>
J <sub>2</sub>	n = 29, omit X <sub>2</sub>
⋮	
J <sub>30</sub>	n = 29, omit X <sub>30</sub>

- will produce similar results from sample to sample (bootstrap may not)

## **Hypothesis Testing**

- a. explain hypothesis testing and its components, including statistical significance, Type I and Type II errors, and the power of a test.
- b. construct hypothesis tests and determine their statistical significance, the associated Type I and Type II errors, and power of the test given a significance level
- c. compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test



- the null ( $H_0$ ) always contains the equality sign

$$H_0: \bar{X} = \mu_0 \quad H_0: \bar{X} \leq \mu_0 \quad H_0: \bar{X} \geq \mu_0$$

- testing  $H_0$  is always done at equality

**Test Statistic:**

(Step #2)

pop.  $\sigma^2$  is known

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

distributed normally

pop.  $\sigma^2$  is unknown

$$t = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

t-distributed

What We Want to Test	Test Statistic	Probability Distribution of the Statistic	Degrees of Freedom
Test of a single mean	$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$	t-Distributed	$n - 1$
Test of the difference in means	$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$	t-Distributed	$n_1 + n_2 - 2$
Test of the mean of differences	$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}}$	t-Distributed	$n - 1$
Test of a single variance	$\chi^2 = \frac{s^2(n-1)}{\sigma_0^2}$	Chi-square distributed	$n - 1$
Test of the difference in variances	$F = \frac{s_1^2}{s_2^2}$	F-distributed	$n_1 - 1, n_2 - 1$
Test of a correlation	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	t-Distributed	$n - 2$
Test of independence (categorical data)	$\chi^2 = \sum_{i=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$	Chi-square distributed	$(r - 1)(c - 1)$

**Step 3: Specify the Level of Significance**

- level of sig. depends on the seriousness of making

a mistake

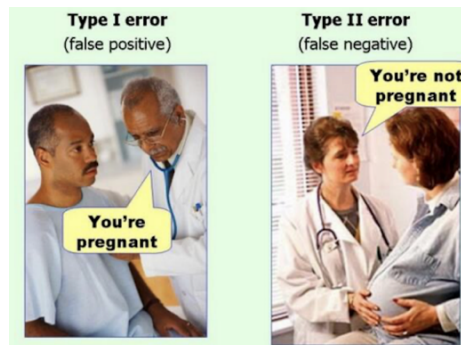
	$H_0 = \text{true}$	$H_0 = \text{false}$
fail to reject	Correct ( $1 - \alpha$ ) confidence level	Type II error $\beta$
reject	Type I error $\alpha$ level of sig.	Correct ( $1 - \beta$ ) Power of a test

• as  $\alpha \downarrow$ ,  $\beta \uparrow$

• only way to decrease both is to increase  $n$

$$t = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \rightarrow \text{as } n \uparrow, \text{ denom. } \downarrow, \text{ t-stat } \uparrow$$

e.g.:  $H_0$ : not pregnant  
 $H_a$ : pregnant

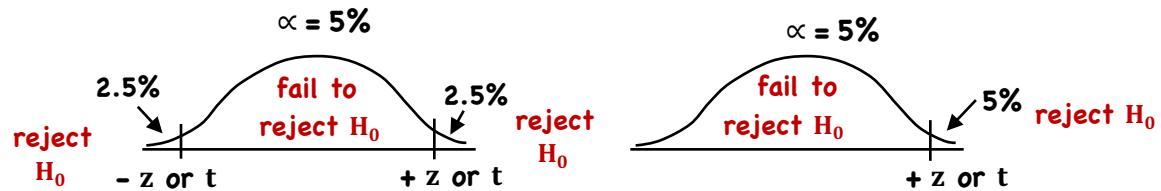


**Step #4: State the Decision Rule**

2-tail: Reject  $H_0$  when  $|test - statistic| > |critical value|$   
 $t$  or  $z$                        $t_{\alpha/2}$  or  $z_{\alpha/2}$

right tail: Reject  $H_0$  when test-statistic  $>$  critical value ( $t_{\alpha}$  or  $z_{\alpha}$ )

left tail : Reject  $H_0$  when test-statistic  $<$  critical value



Cut-off for ...	Excel	Python	R
Right tail, 2.5%	NORM.S.INV(0.975)	norm.ppf(.975)	qnorm(.025,lower.tail=FALSE)
Left tail, 2.5%	NORM.S.INV(0.025)	norm.ppf(.025)	qnorm(.025,lower.tail=TRUE)
Right tail, 5%	NORM.S.INV(0.95)	norm.ppf(.95)	qnorm(.05,lower.tail=FALSE)
Left tail, 5%	NORM.S.INV(0.05)	norm.ppf(.05)	qnorm(.05,lower.tail=TRUE)

or T.INV(p,df)

**Parametric Testing**  
 sample stats to test pop. parameters  
 ( $\bar{X}$  or  $S^2$  for  $\mu$  or  $\sigma^2$ )  
 distributional assumptions

**Non-parametric testing**  
 no parameters tested  
 no distributional assumptions

- when data do not meet distributional assumptions  
 i.e.  $n < 30$ , pop. is non-normal
- when there are outliers  
 - test of median instead of mean
- when data are given in ranks or use an ordinal scale  
 NO ordered categorical IR
- hypothesis do not concern a parameter  
 e.g. Is a sample random?

## **Parametric and Non-Parametric Tests of Independence**

- a. 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
- b. explain tests of independence based on contingency table data

## Hypothesis Testing

Page 1

• **Tests of Correlation**

**1/ Parametric test**

**2-sided**  $H_0: \rho = 0$   
 $H_a: \rho \neq 0$

<b>left</b>	<b>right</b>
<b>one-sided</b> $H_0: \rho \geq 0$	$H_0: \rho \leq 0$
$H_a: \rho < 0$	$H_a: \rho > 0$

- recall  $r_{xy} = \frac{\text{Cov}(x,y)}{S_x S_y}$

→ called a **Pearson correlation**  
or a **Bivariate correlation**

**test-statistic:**  $t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

- in testing  $r$ , as  $n \uparrow$ ,  $H_0$  rejected for even small correlations
- big data sets, almost any  $r$  will be significant

e.g./  $r = .02$   
 $n = 10,000$        $t_{9,998} = \frac{.02\sqrt{9,998}}{\sqrt{1-.02^2}} \sim 2$       vs. **critical  $t = 1.96$**   
**ex. #1**

Page 2

• **Tests of Correlation**

**2/ Non-parametric test**

- if normality assumption for X or Y violated, or outliers are present

- **Spearman rank correlation coefficient**

- basically a correlation, but calculated on rank values and not the values of the observations of X or Y

**1/ Rank all X from largest to smallest**

- assign a rank: 1 = largest ... n = smallest
- for a tie: assign all the same → average rank

e.g. 3 tied for 6<sup>th</sup>       $(6 + 7 + 8)/3 = 7$

- each get a rank of 7

- repeat for Y

• Tests of Correlation/

2/ Non-parametric test

2/ On original data set (pre-ranked):

calculate  $d_i^2 = (\text{rank } X_i - \text{rank } Y_i)^2$

- where X and Y are original paired obs.

3/  $r_s = 1 - \left[ \frac{6(\sum_{i=1}^n d_i^2)}{n(n^2 - 1)} \right]$

↓  
test  $r_s$ , if  $n > 30$

$$t_{n-2} = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}}$$

- white text example
- Example #3

• Tests of Independence/

- test if classification types are independent

e.g./ Are growth stocks equally likely to be any size or are they more likely to be large-cap stocks?

• Tests of Independence/ Contingency Table (2-way)

observed Investment Type	Size Based on Market Capitalization			Total
	Small	Medium	Large	
Value	50	110	343	503
Growth	42	122	202	366
Blend	56	149	520	725
Total	148	381	1,065	1,594

• non-parametric test of indep.

$$\chi^2 = \sum_{i=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad df = (r-1)(c-1) \quad (\text{right-tailed})$$

m = # of cells (3 x 3 = 9)

$O_{ij}$  = observed value in each cell

$E_{ij}$  = expected value in each cell

$$E_{ij} = \frac{(\text{row}_i \text{ total}) \times (\text{column}_j \text{ total})}{\text{Overall total}}$$

$$E_{SV} = (503 \times 148) / 1594 = 46.703$$

$$E_{MG} = (366 \times 381) / 1594 = 87.482$$

$$\therefore \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \text{SV} = \frac{(50 - 46.703)^2}{46.703} = .2328$$

A. Expected Frequency of ETFs by Size and Investment Type

Investment Type	Size Based on Market Capitalization	Small	Medium	Large
		Value $E_{SV} \rightarrow$	46.703	120.228
Growth		33.982	87.482 $\rightarrow E_{MG}$	244.536
Blend		67.315	173.290	484.395
Total		148.000	381.000	1,065.000

B. Scaled Squared Deviation for Each Combination of Size and Investment Type

Investment Type	Size Based on Market Capitalization		
	Small	Medium	Large
Value	0.233	0.870	0.143
Growth	1.892	13.620	7.399
Blend	1.902	3.405	2.617

• Tests of Independence/

$H_0$  : size and type are independent

$H_a$  : size and type are not independent

B. Scaled Squared Deviation for Each Combination of Size and Investment Type

Investment Type	Size Based on Market Capitalization		
	Small	Medium	Large
Value	0.233	0.870	0.143
Growth	1.892	13.620	7.399
Blend	1.902	3.405	2.617

$$\sum_{i=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 32.08025$$

$$= \text{CHISQ.INV}(0.95,4) = 9.4877$$

∴ reject  $H_0$

standardized residuals =  $\frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$

→ > 0 means more obs. than expected if categories were independent

< 0 - opposite (i.e. fewer obs.)

e.g./  $SR_{SV} = \frac{50 - 46.703}{\sqrt{46.703}} = 0.48$

$SR_{MG} = \frac{122 - 87.482}{\sqrt{87.482}} = 3.69$

} more than expected if independent

example #4

## **Simple Linear Regression**

- a. describe a simple linear regression model, how the least squares criterion is used to estimate regression coefficients, and the interpretation of these coefficients
- b. explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated
- c. calculate and interpret measures of fit and formulate and evaluate tests of fit and of regression coefficients in a simple linear regression
- d. 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
- e. calculate and interpret the predicted value for the dependent variable, and a prediction interval for it, given an estimated linear regression model and a value for the independent variable
- f. describe different functional forms of simple linear regressions

## Simple Linear Regression

LOS a (2.5p)	Simple Linear Regression - describe
LOS b (8.5p)	Estimating Parameters - describe
LOS c (7p)	Assumptions of SLR - explain, describe
LOS d, e (5.5p)	Analysis of Variance - calculate, interpret, describe
LOS f (8p)	Hypothesis Testing → Coefficients - formulate, determine
LOS g (4p)	Prediction & Prediction Intervals - calculate, interpret
LOS h (7p)	Functional Forms of LR - describe

Page 1  
LOS a  
- describe

- Simple Linear Regression (LR) → one IV

DV - dependent variable - Y - the variable we are seeking to explain  
 IV - independent variable - X → the explanatory variable

LR assumes a linear relationship between the DV and the IV

Variation of  $Y = \sum_{i=1}^n (Y_i - \bar{Y})^2$  → SST or total sum of squares

- best guess for Y is  $\bar{Y}$ , thus if X gives a more accurate estimate of Y than  $\bar{Y}$ , we say X helps explain Y

---

$Y_i = b_0 + b_1 X_i + \epsilon_i$

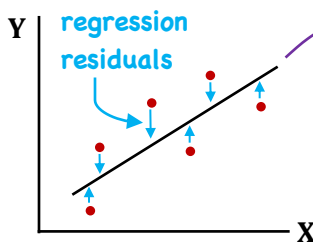
intercept      slope coefficient

regression coefficients

(Y is regressed on X)

$\epsilon$  → error term (residual) - describe

- the portion of the DV that cannot be explained by the IV



$(\bar{X}, \bar{Y})$  lies on the regression line

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$

compute a line of best fit that minimizes the sum of the squared deviations between the observed values of Y and the predicted values (the regression line)

$$\text{i.e. } \min. \sum_{i=1}^n (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2 = \sum_{i=1}^n \varepsilon_i^2 \rightarrow \text{SSE}$$

| DV predicted values of DV ( $\hat{Y}$ )

- sum of the squares error

(a.k.a. residual sum of squares)

$$\begin{cases} E(\varepsilon) = 0 \\ \sum \varepsilon_i = 0 \end{cases}$$

- Note:  $\varepsilon = Y_i - \hat{Y}$  implies the residual is in the same units of measurement as the DV (Y)

$$\hat{b}_1 = \frac{\text{Cov}(X, Y)}{\sigma_x^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

→ denominator can never be negative, ∴ sign of  $\hat{b}_1$  is determined solely by  $\text{Cov}(X, Y)$   
- if  $\hat{b}_1 > 0, r_{XY} > 0$

• Interpreting  $\hat{b}_0$  and  $\hat{b}_1$

$\hat{Y} = b_0$  if  $X_i = 0$  → only makes sense if the IV has meaning at  $X = 0$

$\hat{b}_1$  → the change in Y for a one unit change in X

e.g.  $\text{ROA}(\%) = 4.875\% + 1.25 \cdot \text{CAPEX}(\%)$  → ROA = 4.875% if CAPEX = 0  
→ if CAPEX ↑ 1 unit (i.e. 1%), then ROA ↑ 1.25%

Data/ • cross-sectional - many observations on X & Y for the same time period

• time-series - many observations on Y (and sometimes X) from different time periods

example #2/3

Page 4  
LOS c  
- explain  
- describe

• Assumptions/

1/ **Linearity** → the relationship between X & Y is linear in the parameters  $b_0$  and  $b_1$  → neither is multiplied or divided by another regression parameter

→ implies the IV must not be random - if so, there would be no linear relation between X & Y

2/ **Homoskedasticity** →  $\text{Var}(\varepsilon)$  is the same for all observations (vs. heteroskedastic) - a violation indicates the data series may come from 2 different populations (CS) or regimes (TS)

3/ **Independence** → the pairs (X, Y) are independent of each other  
∴  $\varepsilon$  is uncorrelated across observations (no serial correlation)  
- needed to correctly estimate the variances of  $b_0$  and  $b_1$

Page 5  
LOS c  
- explain  
- describe

• Assumptions/

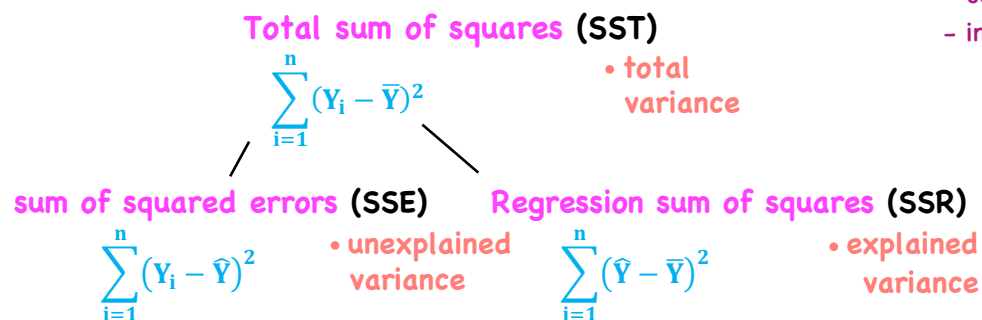
4/ **Normality** →  $\varepsilon$  is normally distributed

- required to conduct valid tests of the values of the regression coefficients

example #4

• Analysis of Variance/

LOS d  
- calculate  
- interpret



∴  $SST = SSE + SSR$

or/ total SS = unexplained SS + explained SS

- calculate
- interpret

• Analysis of Variance/

- **Coefficient of Determination** - measures the fraction of the total variation in the DV that is explained by the IV (goodness of fit measure)

- if only 1 IV, square the correlation between IV and DV

- if multiple IVs: total variation in  $Y = \sum_{i=1}^n (Y_i - \bar{Y})^2$

explained variation in  $Y = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$

$$R^2 = \frac{SSR}{SST} = \frac{\text{explained var.}}{\text{total var.}} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

} measures fit but is not a statistical test

↓  
Coeff. of Determination

∴ statistical test = F-test

$$H_0: b_1 = 0$$

$$H_a: b_1 \neq 0$$

multiple IVs

$$H_0: b_1 = b_2 = \dots = b_k = 0$$

$$H_a: \text{at least one } b_k \neq 0$$

- calculate
- interpret

• Analysis of Variance/

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

df<sub>1</sub> = k  
df<sub>2</sub> = n - k - 1

/ mean

k = slope

coefficients

k + 1 = regression coefficients

• ANOVA table/

- describe
- calculate
- interpret

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Regression	$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{1}$	$F = \frac{MSR}{MSE} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\frac{1}{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}}$
Error	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	n - 2	$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}$	
Total	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	n - 1	$\sqrt{MSE} = SEE$	

- LOS e  
- describe  
- calculate  
- interpret

• **Standard Error of the Estimate (SEE) - a measure of the s.d. of  $\hat{\epsilon}_i$**

$$SEE = \sqrt{MSE} = \left( \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2} \right)^{1/2} = \left( \frac{\sum_{i=1}^n \epsilon_i^2}{n - 2} \right)^{1/2}$$

the smaller the SEE, the more accurate the regression

(a.k.a. the standard error of the regression or the root mean square error)

e.g./

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Regression	191.625	1	191.625	16.0104
Error	47.875	4	11.96875	
Total	239.50	5		

$$= F.INV(.95, 1, 4)$$

$$= 7.71$$

$$\rightarrow SEE = \sqrt{11.96875}$$

**Example #5**

- LOS f  
- formulate  
- determine

1/ Hypothesis Tests of  $\hat{b}_1$  :

test statistic:  $t = \frac{\hat{b}_1 - b_1}{S_{\hat{b}_1}}$   $\leftarrow$  hypothesized value

$$df = n - (k + 1) \quad S_{\hat{b}_1} \rightarrow \text{standard error of } \hat{b}_1 = \frac{SEE}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

e.g./  $b_1 = 1.25 \quad \sum_{i=1}^n (X_i - \bar{X})^2 = 122.64$

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Regression	191.625	1 k	191.625	16.0104
Error	47.875	4 n-(k+1)	11.96875	
Total	239.50	5 n-1		

$$H_0: b_1 = 0 \quad \text{at } \alpha = 5\%$$

$$H_a: b_1 = \neq 0$$

$$= T.INV(.05, 4) = 2.776$$

$$\downarrow$$

$$n - (k + 1)$$

$$S_{\hat{b}_1} = \frac{\sqrt{11.9678}}{\sqrt{122.64}} = 0.312398 \rightarrow t = \frac{1.25 - 0}{.312398} = 4.00131 \rightarrow \text{Reject } H_0$$

( $t^2 = F, 4.00131^2 = 16.0104$ )

Note:  $H_0: \rho = 0$

$H_a: \rho \neq 0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \rightarrow \frac{0.8945\sqrt{4}}{\sqrt{1-.8945^2}} = 4.00131 \rightarrow \text{Reject } H_0$$

$$df = n - 2$$

SLR only





3/ Log-Log model  $\ln Y_i = b_0 + b_1 \cdot \ln(X_i)$

↓  
the relative change in  $Y_i$  for a  
relative change in  $X_i$

- exh. #37/38

• selecting a model depends on goodness of fit

- $R^2$
- F-stat
- SEE ( $S_c$ )

- a plot of the residuals should show randomness and the distribution should be normal

- if not, consider transforming the DV, IV, or both.

## **Introduction to Big Data Techniques**

- a. describe aspects of “fintech” that are directly relevant for the gathering and analyzing of financial data.
- b. describe Big Data, artificial intelligence, and machine learning
- c. describe applications of Big Data and Data Science to investment management

## Intro. to Big Data Techniques

Page 1

**Fintech - technological innovation in the design and delivery of financial services and products**

- **analysis of large datasets**
  - traditional data
  - alternative data - social media, sensor networks
- **analytical tools** - for extremely large datasets → AI → very useful for complex non-linear relationships

**Big Data/ data generated from traditional/alternative sources**

### Characteristics

- quality vs. quantity
- **Volume** - millions to billions of data points
  - **Velocity** - speed of recording/transmitting/generating has accelerated (real time/near real-time)
  - **Variety** - text, image, speech, video
  - **Veracity** - credibility/reliability of different sources

Page 2

- semi (
- **structured** - can be organized in tables, database entries
  - **unstructured** - cannot be represented in tabular format (voice, images) (all data must be structured before use)

### Sources/

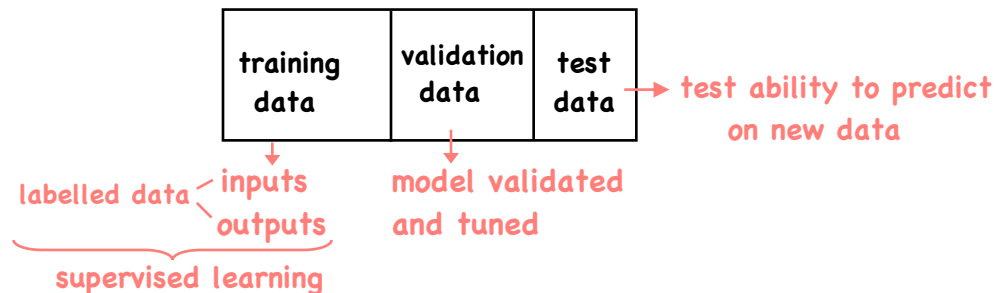
- **financial markets** - price/volume
- **businesses** - filings, press releases
- **governments** - economic data
- **individuals** - web footprints, transactional data
- **sensors** - images, traffic volume
- **Internet of Things** - smart buildings, appliances

### 3 main sources of alternative data:

- **Individuals** (text, photos, audio, website clicks/engagement)
- **Business processes** (sales, credit card data, corporate exhaust)
  - **supply chain info.**
  - **PoS scanner data etc...**
- **Sensor data** (smartphones, RFID chips)
  - **IoT - array of physical devices**

- Ethical/Legal issues → use of personal info./data scraped from web data
- Challenges
  - quality
  - volume
  - appropriateness
- selection bias
- outliers
- missing data
- data cleansing/organizing
- AI - artificial intelligence - systems that exhibit cognitive and decision-making ability comparable or superior to humans
  - from expert systems to neural networks
    - (if-then rules)
    - (based on how brains learn and process info.)
- Machine learning - seek to extract knowledge from large amounts of data without assumptions of the data's underlying probability distribution
  - learn from known examples to determine underlying structure in the data - generate structure without help from a human

- requires large training datasets



- over/under-fit - will not perform well out-of-sample
- Unsupervised learning - only inputs, no outputs
- Deep learning - multi-stage, non-linear data processing (supervised or unsupervised)
- Data Science - comp. science + statistics to extract info. from big data

### Data processing methods/

- **Capture** - low latency systems (very little delay) vs. high latency
- **Curation** - ensuring data quality/accuracy
  - detect errors and make adjustments for missing data
- **Storage** - record/archive/access in databases
- **Search** - how data are queried
- **Transfer** - movement of data from storage to analytical application

### Data Visualization/

- how data is formatted/displayed/summarized in graphical format
  - (e.g. heat maps, tree diagrams, network graphs)
- tag clouds (text data) - words are sized/displayed based on frequency

- **Text analytics** - use large unstructured voice or text datasets
- **Natural language processing (NLP)** - comp. science + AI + linguistics
  - analyze/interpret human language
    - (e.g. translation, speech recognition)