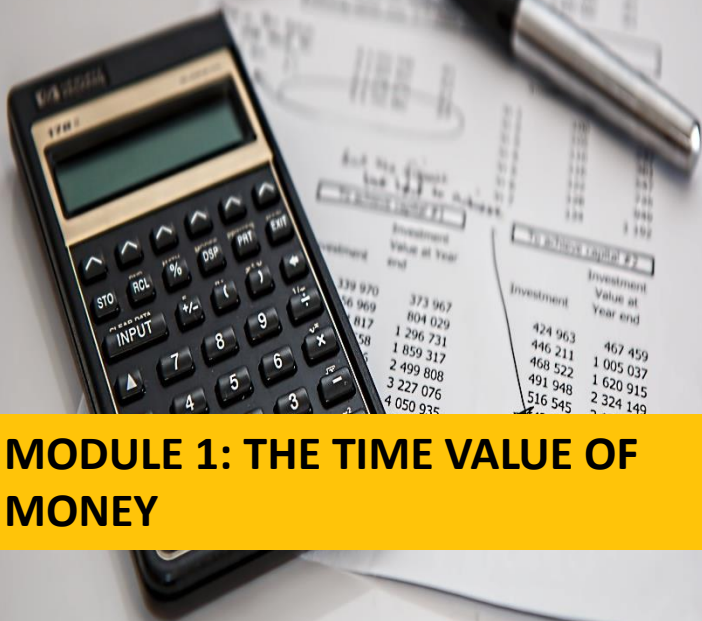




2023
LEVEL 1

QUANTITATIVE METHODS

**LECTURE SLIDES FOR
CFA PROGRAM LEVEL 1 2023**



MODULE 1: THE TIME VALUE OF MONEY

MODULE 1: THE TIME VALUE OF MONEY

Learning outcome statements

Interest rates, Future value, Present value

[LOS 1.a] Interpret interest rates as required rates of return, discount rates, or opportunity costs

[LOS 1.b] Explain an interest rate as the sum of a real risk-free rate and premiums that compensate investors for bearing distinct types of risk

[LOS 1.e] Calculate and interpret the future value (FV) and present value (PV) of a single sum of money, an ordinary annuity, a perpetuity (PV only), and a series of unequal cash flows

[LOS 1.f] Demonstrate the use of a timeline in modeling and solving time value of money problems

Compounding frequency and effective annual rate

[LOS 1.c] Calculate and interpret the effective annual rate, given the stated annual interest rate and the frequency of compounding

[LOS 1.d] Calculate the solution for time value of money problem with different frequencies of compounding

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.a] Interpret interest rates as required rates of return, discount rates or opportunity cost

1.

Definition of interest rate

The interest rate is the amount a lender charges for the use of assets expressed as a percentage of the principal.

2.

Interpretation of interest rate

Discount rate

The rate at which investors discount cashflows in the future to present value.

(Refer to example 1)

Opportunity cost

The value that investors forgo by choosing a course of action.

(Refer to example 2)

Required rate of return

The minimum rate of return an investor must receive in order to accept the investment.

Example 1: Discount rate

If an individual can borrow funds at an interest rate of 10%, then that individual should discount payments in the future at 10% in order to get their equivalent value.

Example 2: Opportunity cost

The interest rate on Hoa Phat's bond is 8%, which is the value that investors forgo when investing in Hoa Phat's stock → the opportunity cost of investing in Hoa Phat's stock is 8%.

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.b] Explain interest rate as the sum of a real risk free rate and premiums

3.

Relationship equation

Example: Illustrating the relationship between required rate of return and risk-free rate

Investor A lends his relative \$1 million dollar. The return of US. Treasury bill (risk-free rate) is 5%.

Comment:

When A lends his relative \$1 million, he has to bear the risk that the relative will default, and he can't get the payment for the amount of money he lent.

→ For that reason, A wants 2% more of the risk-free rate to compensate for the risk that he has to bear when lending money to the relative.

We can call 2% as “a default risk premium.”

→ The total return that A will get for the loan is:

Risk-free rate + default risk premium = $5\% + 2\% = 7\%$.

Beside default risk premium, we have 3 other types of premium, together they are added to the risk free rate to compute the required rate of return, explained below.

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.b] Explain interest rate as the sum of a real risk free rate and premiums

3.

Relationship equation

Required rate of return

Nominal risk free rate

Nominal risk free rate = Real risk free rate + Expected inflation.

Note: Real risk free rate is mainly theoretical

Default risk premium

Excess return compensates for the risk that a borrower will not make the promised payments in a timely manner.

Liquidity risk premium

Excess return compensates for the risk of receiving less than fair value for an investment if it must be sold for cash quickly.

Maturity risk premium

Excess return compensates investors for the increased sensitivity of the market value of debt to a change in market interest rates as maturity is extended.

MODULE 1: THE TIME VALUE OF MONEY

Basic knowledge: Simple interest vs compound interest

Simple interest

- Simple interest is based on the principal amount of a loan or deposit.
- Simple interest involves no reinvestment of the interest we receive.

Compound interest

- Compound interest is based on the principal amount and the interest rate that accumulates on it every period.
- Compound interest involves reinvestment of the interest we receive.

Analyzing the value of a \$100 4-year investment in two cases: (1) simple interest rate of 10% annually and (2) compound interest rate of 10% annually

1. Simple interest: no reinvestment

	t = 0	Year 1	Year 2	Year 3	Year 4
Amount to calculate interest		100	100	100	100
Interest received		$100 \times 10\% = 10$	10	10	10
Total value of the investment		$100 + 10 = 110$	$110 + 10 = 120$	$120 + 10 = 130$	$130 + 10 = 140$

2. Compound interest: reinvestment involved

	t = 0	Year 1	Year 2	Year 3	Year 4
Amount to calculate interest		100	110	121	133.1
Interest received		$100 \times 10\% = 10$	$110 \times 10\% = 11$	$121 \times 10\% = 12.1$	13.31
Total value of the investment		$100 + 10 = 110$	$110 + 11 = 121$	$121 + 12.1 = 133.1$	$133.1 + 13.31 = 146.41$

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret future value and present value of a lump-sum

Formula	Definition	Future value	Present value
		The amount to which a current deposit will grow over time when it is placed in an account paying compound interest	Today's value of a cash flow that is to be received at some point in the future
Annually compounding		$FV = PV \times (1 + r)^N$ <i>r = stated annual rate</i> <i>N = number of years</i>	$PV = \frac{FV}{(1 + r)^N}$
Periodically compounding (*)		$FV = PV \times \left(1 + \frac{r}{m}\right)^{N \times m}$ <i>m = the number of compounding periods per year</i> <i>r = stated annual rate</i> <i>N = number of years</i>	$PV = \frac{FV}{\left(1 + \frac{r}{m}\right)^{N \times m}}$
Continuously compounding (**)		$FV = PV \times e^{r \times N}$ <i>r = stated annual rate</i> <i>N = number of years</i>	$PV = \frac{FV}{e^{r \times N}}$

(*) The formula of periodically compounding can also be written as:

$$FV = PV \times (1 + \text{periodic interest rate})^{\text{Number of periods}}$$

$$PV = \frac{FV}{(1 + \text{periodic interest rate})^{\text{Number of periods}}}$$

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret future value and present value of a lump-sum

(**) Justifying the formula of continuous compounding

- Continuous compounding means that the amount of money at $t = 0$ is compounded for an indefinite number of periods ($m \rightarrow +\infty$)
- If we want to use the future value formula with continuous compounding, we need to find the limiting value of the future value factor for $m \rightarrow +\infty$.

[Additional reading] Prove the continuous compounded-FV formula

$$FV_{\text{continuous}} = \lim_{m \rightarrow +\infty} PV \times \left(1 + \frac{r}{m}\right)^{N \times m}$$

$$y = \frac{m}{r} \rightarrow m = y \times r \text{ then we have: } PV \times \left(1 + \frac{r}{m}\right)^{N \times m} = PV \times \left(1 + \frac{1}{y}\right)^{N \times y \times r}$$

$$\rightarrow FV_{\text{continuous}} = \lim_{m \rightarrow +\infty} PV \times \left(1 + \frac{r}{m}\right)^{N \times m} = \lim_{y \rightarrow +\infty} PV \times \left(1 + \frac{1}{y}\right)^{N \times y \times r}$$

$$= PV \times \lim_{y \rightarrow +\infty} \left(1 + \frac{1}{y}\right)^{N \times y \times r}$$

$$= PV \times \lim_{y \rightarrow +\infty} \left[\left(1 + \frac{1}{y}\right)^y \right]^{N \times r} = PV \times \left[\lim_{y \rightarrow +\infty} \left(1 + \frac{1}{y}\right)^y \right]^{N \times r} = PV \times e^{r \times N}$$

$$(\text{remember that } \lim_{y \rightarrow +\infty} \left(1 + \frac{1}{y}\right)^y = e \approx 2.718)$$

$$FV = PV \times e^{r \times N}$$

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret future value and present value of a lump-sum

Implication about the frequency of compounding

The table below shows how a stated annual interest rate of 8% generates different future value with annual, semiannual, quarterly, monthly, daily, and continuous compounding for an initial investment of \$1.

Frequency	r/m	mN	Future Value of \$1		
Annual	$8\%/1 = 8\%$	1	$\$1.00 \times (1.08)$	=	\$1.08
Semiannual	$8\%/2 = 4\%$	2	$\$1.00 \times 1.04^2$	=	\$1.081600
Quarterly	$8\%/4 = 2\%$	4	$\$1.00 \times 1.02^4$	=	\$1.082432
Monthly	$8\%/12 = 0.6667\%$	12	$\$1.00 \times 1.006667^{12}$	=	\$1.083000
Daily	$8\%/365 = 0.0219\%$	365	$\$1.00 \times 1.000219^{365}$	=	\$1.083278
Continuous			$\$1.00 \times e^{[0.08 \times 1]}$	=	\$1.083287

FV increases as frequency increases

Limiting value

$FV_{\text{annually cpd}} < FV_{\text{periodically cpd}} < FV_{\text{continuously cpd}}$

→ We will receive better amount of future value as we increase the frequency of compounding, and we get the maximum value when the deposit is continuously compounded.

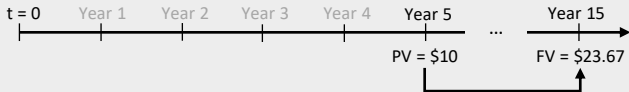
MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret future value and present value of a lump-sum

Example: Calculating the future value and present value

1. A pension fund manager estimates that his corporate sponsor will make a \$10 million contribution five years from now. The rate of return on plan assets has been estimated at 9 percent per year. The pension fund manager wants to calculate the future value of this contribution 15 years from now, which is the date at which the funds will be distributed to retirees. What is that future value?
2. Given a discount rate of 10%, calculate the PV of a \$200 cash flow that will be received in two years. Interests are compounded quarterly.

Answer:



Using a non-financial calculator:

1. This is the case of annual compounding, so $FV = PV \times (1 + r)^N$
 And $N = 10$, $r = 9\%$

$$\rightarrow FV_{10} = PV \times (1 + r)^N = 10 \times (1 + 9\%)^{10} = 23.67$$

2. This is the case of periodically compounding, so $PV = FV / \left(1 + \frac{r}{m}\right)^{N \times m}$

And $m = 4$ (compounded quarterly), $N = 2$

$$\rightarrow PV = FV_2 / \left(1 + \frac{r}{m}\right)^{N \times m} = 200 / \left(1 + \frac{10\%}{4}\right)^{4 \times 2} = 164.15$$

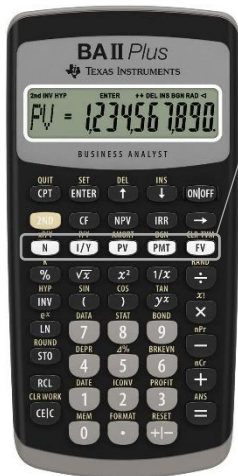
MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret future value and present value of a lump-sum

Example: Calculating the future value and present value

Answer:

Using a financial calculator:



Use these functions to calculate number of periods, discount rate, present value, payment or future value when the other components are known.

N

Enter or calculate the number of periods

I/Y

Enter or calculate the discount rate

PV

Enter the cashflow at $t = 0$ or calculate the present value

PMT

Enter or calculate the payment each year

FV

Enter the cash flow at the final period or calculate the future value

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret future value and present value of a lump-sum

Example: Calculating the future value and present value

Answer:





Using a financial calculator:

1.



- N = the number of periods = 10
- r = discount rate = 9%
- PV = initial cashflow ($t = 0$) = \$10
- PMT = payment at period = 0

→ This is how we use the calculator:

Step 1: Enter the components

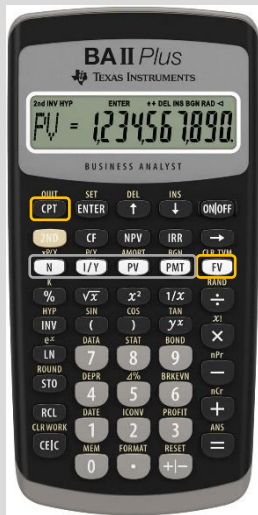
- Press 10, then press 
- Press 9, then press 
- Press -10, then press 
- Press 0, then press 

Step 2: Calculate FV


- Press , then press 

The result shows $FV = 23.67$

2. Similar to the steps at exercise 1.



 Inputs

 Output

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret FV and PV of an ordinary annuity and an annuity due

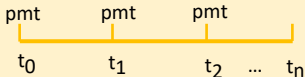
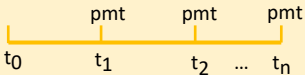
1.

Definition of ordinary annuity and annuity due

An annuity is a finite set of equal cash flows that occurs at equal intervals

An **ordinary annuity** has a first cash flow (PMT) that occurs one period from now (t_1).

An **annuity due** has a first cash flow (PMT) that occurs immediately (t_0).



→ The cash flows of an annuity due is **ahead** those of an ordinary annuity by **1 time period**.

(*) *PMT = the equal cash flow that occurs periodically*

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret FV and PV of an ordinary annuity and an annuity due

2.

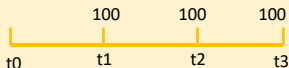
Calculation relating to ordinary annuity and annuity due

Ordinary annuity

Formula:

- $$FV_{\text{ordinary}} = PMT \times \left[\frac{(1+r)^N - 1}{r} \right]$$
- $$PV_{\text{ordinary}} = PMT \times \left[\frac{1 - (1+r)^{-N}}{r} \right]$$

Example 1: (given $r = 10\%$)



$$FV_{\text{ordinary}} = 100 \times \left[\frac{(1 + 10\%)^3 - 1}{10\%} \right]$$

$$= 331$$

Use financial calculator:

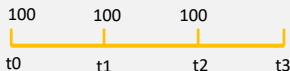
$N = 3$, $I/Y = 10$, $PMT(END) = 100$,
 $PV = 0 \rightarrow FV = -331$

Annuity due

Formula:

- $$FV_{\text{due}} = FV_{\text{ordinary}} \times (1+r)$$
- $$PV_{\text{due}} = PV_{\text{ordinary}} \times (1+r)$$

Example 1: (given $r = 10\%$)



$$FV_{\text{due}} = 100 \times \left[\frac{(1+10\%)^3 - 1}{10\%} \right] \times (1+10\%)$$

$$= 364.1$$

Use financial calculator:

$N = 3$, $I/Y = 10$, $PMT(END) = 100$, $PV = 0 \rightarrow FV = -331$

$\rightarrow FV_{\text{due}} = 331 \times (1 + 10\%) = 364.1$

Refer to slide 12-13 to see how to use the calculator.

MODULE 1: THE TIME VALUE OF MONEY

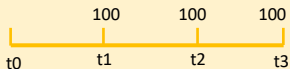
[LOS 1.e] Calculate and interpret FV and PV of an ordinary annuity and an annuity due

2.

Calculation relating to ordinary annuity and annuity due

Ordinary annuity

Example 2: (given $r = 10\%$)



$$PV_{\text{ordinary}} = 100 \times \left[\frac{1 - (1 + 10\%)^{-3}}{10\%} \right]$$

$$= 248.7$$

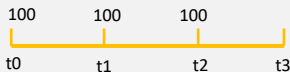
Use financial calculator:

$N = 3$, $I/Y = 10$, $PMT(END) = 100$, $FV = 0$

→ $PV = -248.7$

Annuity due

Example 2: (given $r = 10\%$)



$$PV_{\text{due}} = 100 \times \left[\frac{1 - (1 + 10\%)^{-3}}{10\%} \right] \times (1 + 10\%)$$

$$= 273.6$$

Use financial calculator:

$N = 3$, $I/Y = 10$, $PMT(END) = 100$, $FV = 0$

→ $PV = -248.7$

→ $PV_{\text{due}} = 248.7 \times (1 + 10\%) = 273.6$

Refer to slide 12-13 to see how to use the calculator.

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret PV of a perpetuity

Perpetuity

Definition: A perpetuity is a **perpetual annuity**, or a set of level **never-ending** sequential cash flows, with the first cash flow occurring one period from now.

Formula:

$$PV_0 = \frac{PMT_1}{r}$$

PMT_1 = the payment you will receive next year

Note: In case of a deferred perpetuity, if the payment a perpetuity is deferred until year n , we must discount PV by $\frac{1}{(1+r)^{n-1}}$
(refer to example below)

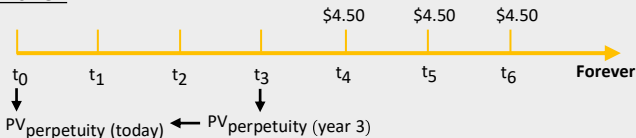
Example:

Kodon Corporation issues preferred stock that will pay \$4.50 per year in annual dividends beginning 4 years from now and plans to follow this dividend policy forever. Given an 8% rate of return, what is the value of Kodon's preferred stock today?

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret PV of a perpetuity

Answer:



Step 1: Calculating the present value of the perpetuity in year 3

$$PV_{\text{perpetuity (year 3)}} = \frac{PMT_{\text{in year 4}}}{r} = \frac{4.50}{8\%} = 56.25$$

Step 2: Discount the present value of the perpetuity in year 3 to today ($t=0$)

$$PV_{\text{perpetuity (today)}} = \frac{PV_{\text{perpetuity (year 3)}}}{(1+r)^4 - 1} = \frac{56.25}{(1+8\%)^3} = 44.65$$

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret FV and PV of an unequal cash flows

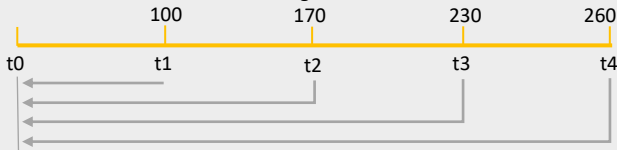
Calculating the present value (PV) or future value (FV) of a series of unequal cash flows can be done by summing the PV or FV of each individual cash flow:

$$\bullet PV_0 = \frac{CF_1}{(1+r)^1} + \frac{CF_2}{(1+r)^2} + \dots + \frac{CF_n}{(1+r)^n}$$

$$\bullet FV_0 = CF_1(1+r)^{n-1} + CF_2(1+r)^{n-2} + \dots + CF_n$$

Illustration: Calculate the present value of an uneven cashflow

Find the PV of the cash flows below given $r = 10\%$



$$PV \text{ of } CF_1 = 100 / (1.1^1) = 90.9$$

$$PV \text{ of } CF_2 = 170 / (1.1^2) = 140.5$$

$$PV \text{ of } CF_3 = 230 / (1.1^3) = 172.8$$

$$PV \text{ of } CF_4 = 260 / (1.1^4) = 177.6$$

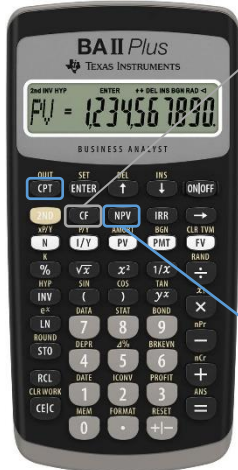
PV of this series of cash flows = 581.7

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret FV and PV of an unequal cash flows

Example: Calculate the present value of an uneven cashflow

Use a financial calculator:



Use this function to enter the cashflows at each period

What would be shown on the screen:

CF₀: Enter the cashflow at $t = 0$

C₀₁: Enter the cashflow at $t = 1$

F₀₁: Enter the frequency of C_{01} , for example, if the cashflow C_{01} also appears at $t = 2$, we can say C_{01} appear twice and we enter $F_{01} = 2$

...

C_{0n}: Enter the cashflow at $t=n$

F_{0n}: Enter the frequency of C_{0n}

Press **CPT** & **NPV** to calculate the present value of cash flows

What would be shown on the screen:

I: Enter the discount rate.

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.e] Calculate and interpret FV and PV of an unequal cash flows

Example: Calculate the present value of an uneven cashflow

Use a financial calculator:

Step 1: Entering the cashflows

Action	Display
CF	CF0
0 ENTER ↓	C01
100 ENTER ↓	F01
1 ENTER ↓	C02
170 ENTER ↓	F02
1 ENTER ↓	C03
230 ENTER ↓	F03
1 ENTER ↓	C04
260 ENTER ↓	F04
1 ENTER ↓	



MODULE 1: THE TIME VALUE OF MONEY

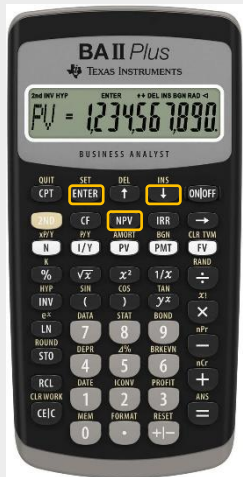
[LOS 1.e] Calculate and interpret FV and PV of an unequal cash flows

Example: Calculate the present value of an uneven cashflow

Use a financial calculator:

Step 2: Enter the discount rate and calculate the cashflow

- Press **NPV**
- The screen displays “I =”
- Enter the discount rate 10 **ENTER** **↓**
- Press **CPT** to show the result:
NPV = 581.7909



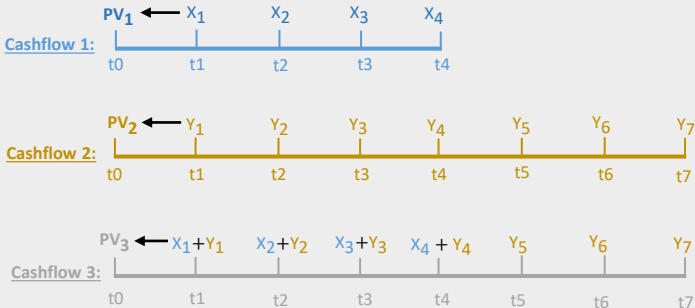
MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.f] Demonstrate the use of a time line in modelling and solving time value of money

Cash flow additivity rule

Key principle: Several amounts of money indexed at the same point in time can be added to form a single cash flow (*Refer to the illustration below*).

Usage: The cash flow additivity principle can be used to solve problems with uneven cash flows by combining single payments and annuities.



$$\text{Cashflow 3} = \text{Cashflow 1} + \text{Cashflow 2} \rightarrow PV_3 = PV_1 + PV_2$$

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.f] Demonstrate the use of a time line in modelling and solving time value of money

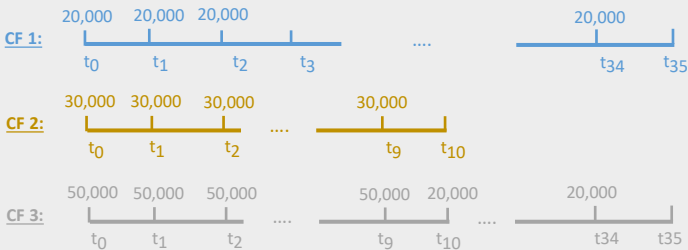
Example: Illustration of cash flow additivity rule

Question: A client plans to retire in 15 years and will need to withdraw \$50,000 from his retirement account each year for 10 years, beginning on the day he retires. After that, he will need to withdraw \$20,000 per year for 25 years. The account returns 4% annually. The amount he needs to have in the account on the day he retires is closest to:

A) \$580,000. B) \$640,000. C) \$655,000.

Answer:

Step 1: Divide this cashflow into 2 cashflows and calculate FV of each cash flow



MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.f] Demonstrate the use of a time line in modelling and solving time value of money

Example: Illustration of cash flow additivity rule

Answer:

Step 2: Calculate the PV of each cashflow (cashflow 1 & cashflow 2)

Cashflow 1: A 35 year annuity due of \$20,000

Use a financial calculator:

$N = 35$, $PMT = 20,000$, $I/Y = 4$, $FV = 0$;

$CPT PV_{\text{ordinary}} = -373,292 \rightarrow PV_{\text{due}} = -388,224 \rightarrow PV_1 = 388,224$

Cashflow 2: A 10 year annuity due of \$30,000

Use a financial calculator:

$N = 10$, $PMT = 30,000$, $I/Y = 4$, $FV = 0$;

$CPT PV_{\text{ordinary}} = -243,326 \rightarrow PV_{\text{due}} = -253,060 \rightarrow PV_2 = 253,060$

Step 3: Add the PV of two cashflows

$PV = PV_1 + PV_2 = 388,224 + 253,060 = 641,284$

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.c] Calculate and interpret the effective annual rate, given the stated annual rate and compounding frequency

1.

Definition of effective annual rate (EAR)

Effective annual rate (EAR) or effective annual yield (EAY) is the rate investors actually realize as a result of compounding while stated annual rates are rates quoted officially by financial institution without compounding.

2.

Construction of the EAR formula

• The case of periodically compounding

EAR = The rate investors actually realize as a result of compounding

→	The value of A lump sum (A) that is compounded at EAR (%) annually in one year.	=	The value of a lump sum (A) that is compounded at r_p (%) in m periods in one year.
---	---	---	---

→	$A \times (1 + \text{EAR})$	=	$A \times (1 + r_p)^m$
---	-----------------------------	---	------------------------

→	$1 + \text{EAR} = (1 + r_p)^m \rightarrow \text{EAR} = (1 + r_p)^m - 1$		
---	---	--	--

• The case of continuously compounding

Using the same approach, EAR in the case of continuously compounding = $e^{rs} - 1$

(r_p is also known as periodic rate)

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.c] Calculate and interpret the effective annual rate, given the stated annual rate and compounding frequency

3.

EAR formula

Periodic compounding

$$\text{EAR}_{\text{periodic}} = (1 + \text{periodic rate})^m - 1$$

where:

periodic rate = stated annual rate / $m = r_s / m$

m = the number of compounding periods per year

Continuous compounding

$$\text{EAR}_{\text{continuous}} = e^{r_s} - 1$$

where:

$e \approx 2.71$

r_s : annual stated rate

As mentioned in LOS 1.e:

$$\text{FV}_{\text{annually cpd}} < \text{FV}_{\text{periodically cpd}} < \text{FV}_{\text{continuously cpd}}$$

→ The rate investors actually realize as a result of compounding – EAR increases if we increase the frequency of compounding, and it becomes maximum if we compound the initial investment continuously.

→ For a certain stated annual rate, we have the following relationship between EAR and frequency of compounding.

Frequency of compounding (m)

Effective annual rate (EAR)

→ + ∞

$$\text{EAR}_{\text{continuous}} = e^{r_s} - 1$$

MODULE 1: THE TIME VALUE OF MONEY

[LOS 1.d] Calculate the solution for time value of money problem with different frequencies of compounding

As we have learnt about effective annual rate (EAR), now we have two ways to compute FV and PV that is compounded periodically.

Example: John plans to invest \$2,500 in an account that will earn 8% per year with quarterly compounding. How much will be in the account at the end of two years?

Approach 1: Compounding the periodic interest rate

Number of compounding periods: $N \times m = 2 \times 4 = 8$

Periodic rate: $r_s/m = 8\%/4 = 2\%$

→ The 2% periodic rate is compounded 8 times, so the future value of the

investment is: $FV = PV \times \left(1 + \frac{r}{m}\right)^{N \times m} = 2,500 \times (1 + 2\%)^8 \approx \$2,929.15$

Approach 2: Using the EAR

$EAR = (1 + \text{periodic rate})^m - 1 = (1 + 2\%)^4 - 1 \approx 8.24\%$

→ The investment grows at 8.24% in 2 years, so the future value of the

investment is: $FV = PV \times (1 + EAR)^N = 2,500 \times (1 + 8.24\%)^2 \approx \$2,929.15$

→ **Compounding the periodic interest rate for $N \times m$ periods, and compounding the annual effective rate (EAR) for N years brings us the same FV result.**

MODULE 1: THE TIME VALUE OF MONEY

Practicing with a financial calculator

Computing the Number of Periods in an Annuity

Question: Jack needs to accumulate at least \$1,000 with annual deposits of \$80 into his bank account. If the annual interest rate is 10%, how many end-of-year payments are required?

Answer:

We have $FV = \$1,000$, $PMT = \$80$ and the periodic discount rate $(I/Y) = 10\%$

→ We can calculate the number of periodic payments required (N) as:

$FV = \$1,000$; $PMT = -\$80$; $I/Y = 10$; $PV = 0$; $CPT N \rightarrow N = 8.51$ years. (*)

Therefore, 9 deposits are needed to accumulate more than \$1,000 in the bank account with annual deposits of \$80.

Computing the Discount Rate for an Annuity

Question: What rate of return will we earn on an ordinary annuity that requires a \$900 deposit today and promises to pay \$150 at the end of every year for the next 10 years?

Answer:

The amount we need to pay today = $PV = -\$900$

The annual payment we will receive = $PMT = \$150$

The number of payments we will receive = $N = 10$

We will not receive anything beyond the 10 annual payments so $FV = 0$

→ We can calculate the periodic interest rate as:

$PV = -\$900$; $PMT = \$150$; $N = 10$; $FV = 0$; $CPT I/Y \rightarrow I/Y = 10.56\%$ (*)

(*) Refer to slide 12-13 for the usage of N, I/Y, PV, PMT, FV functions.

MODULE 1: THE TIME VALUE OF MONEY

Practicing with a financial calculator

Computing the Fixed Monthly Payment on a Loan

Question: Robert has taken a loan of \$10,000 that he will repay through 15 equal yearly installments. Given an annual interest rate of 7%, how much must he pay every year?

Answer:

The amount borrowed = $PV = \$10,000$

The number of compounding periods = $N = 15$

The periodic interest rate (I/Y) = 7%

→ The annual payment assuming that the loan will be paid off ($FV = 0$) in 15 years is calculated as:

$N = 15$; $I/Y = 7$; $PV = \$10,000$; $FV = 0$; CPT PMT → $PMT = -\$1,097.95$ (*)

(*) Refer to slide 12-13 for the usage of N , I/Y , PV , PMT , FV functions.



MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

Learning outcome statements

[LOS 2.a] Identify and compare data types

[LOS 2.b] Describe how data are organized for quantitative analysis

[LOS 2.c] Interpret frequency and related distributions

[LOS 2.d] Interpret a contingency table

[LOS 2.e] Describe ways that data may be visualized and evaluate uses of specific visualizations

[LOS 2.f] Describe how to select among visualization types

[LOS 2.g] Calculate and interpret measures of central tendency

[LOS 2.h] Evaluate alternative definitions of mean to address an investment problem

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

Learning outcome statements

[LOS 2.i] Calculate quantiles and interpret related visualizations

[LOS 2.j] Calculate and interpret measures of dispersion

[LOS 2.k] Calculate and interpret target downside deviation

[LOS 2.l] Interpret skewness

[LOS 2.m] Interpret kurtosis

[LOS 2.n] Interpret correlation between two variables

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.a] Identify and compare data types

Definition

Data can be defined as a collection of numbers, characters, words, and text—as well as images, audio, and video—in a raw or organized format to represent facts or information.

Data classification

1. Based on statistical perspective

Numerical data

Discrete data

Continuous data

Categorical data

Nominal data

Ordinal data

2. Based on how they are collected

Time series data

Cross sectional data

Panel data

3. Based on their organized form

Structure data

Unstructured data

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.a] Identify and compare data types

1.

Classification of data based on statistical perspective

Numerical data (Quantitative data)

Discrete data

Numerical values that result from a counting process and therefore are *limited* to a finite number of values.

Example: The number of shares on a stock exchange.

Continuous data

Data that can be measured and can take on any numerical value in a specified range of values and therefore is *unlimited* to a finite number of value.

Example: The amount of rainfall in summer.

Categorical data (Qualitative data)

Nominal data

Categorical values that are *not* amenable to being organized in a *logical order*.

Example: Classification of publicly listed stocks in number of sectors, such as bank, real estate, financials and so on.

Ordinal data

Categorical values that can be *logically ordered or ranked*.

Example: Assigning the number 1 to the 100 best performing stocks, the number 2 to the next 100 best performing stocks,... for 1000 small cap growth stock.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.a] Identify and compare data types

2.

Classification of data based on how they are collected

Time series data

Observations taken over a *period of time* at specific and equally spaced time intervals.

Example:

The set of monthly returns on FPT stock from January 2017 to January 2022.

Cross-sectional data

A sample of observations taken at a *single point in time*.

Example:

The returns of FPT stock, MWG stock, and other stocks as of January 1, 2022.

Panel data

Observations over time of the *same characteristic* for *multiple entities*.

Example:

Debt/equity ratios for 20 companies over the most recent 24 quarters.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.a] Identify and compare data types

3.

Classification of data based on organized forms

Structured data

Data that are highly organized in a *pre-defined manner*, usually with repeating pattern.

Example:

- *Market data: data issued by stock exchanges (stock prices,...)*
- *Fundamental data: data contained in financial statements (earnings per share...)*
- *Analytical data: data derived from analytics (cash flow projections...)*

Unstructured data

Data that do *not follow any conventionally organized* forms, typically collected from unconventional sources.

Example:

- *Data produced by individuals (social media posts, web searcher,...).*
- *Data generated by business processes (credit card transactions, corporate regulatory filings,...).*
- *Data generated by sensors (traffic cameras, satellites,...).*

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.b] Describe how data are organized for quantitative analysis

Preparing data for quantitative analysis

Raw data

Raw data can be converted into formatted data by either:

- **One dimensional array (*)**
- **Two dimensional array (*)**

Formatted data

Only formatted data are suitable for quantitative analysis

Quantitative analysis

() Presented in next slide*

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.b] Describe how data are organized for quantitative analysis

	One dimensional array	Two dimensional array (Also known as table)																				
Basic function	A one-dimensional array is a structure of components that can be accessed individually by specifying the position of a component with a single index value .	Two dimensional arrays are indexed by two subscripts , one for the row and one for the column.																				
Uses	Suitable for representing a single variable.	Suitable to represent multiple variables and observations.																				
Example	<table><tr><th>Observation by Day</th><th>Stock price (\$)</th></tr><tr><td>1</td><td>57.21</td></tr><tr><td>2</td><td>58.26</td></tr><tr><td>3</td><td>58.64</td></tr></table> <div><div>↓</div>Position of component</div> <div><div>↓</div>Index value</div>	Observation by Day	Stock price (\$)	1	57.21	2	58.26	3	58.64	<table><tr><th>Year</th><th>Year 1</th><th>Year 2</th></tr><tr><td>Revenue</td><td>3,784</td><td>4,097</td></tr><tr><td>EPS</td><td>1.37</td><td>-0.34</td></tr><tr><td>DPS</td><td>N/A</td><td>N/A</td></tr></table> <div><div>↓</div>Subscripts</div>	Year	Year 1	Year 2	Revenue	3,784	4,097	EPS	1.37	-0.34	DPS	N/A	N/A
Observation by Day	Stock price (\$)																					
1	57.21																					
2	58.26																					
3	58.64																					
Year	Year 1	Year 2																				
Revenue	3,784	4,097																				
EPS	1.37	-0.34																				
DPS	N/A	N/A																				

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.b] Describe how data are organized for quantitative analysis

Example: Organizing raw data into a data table

Suppose you are conducting a valuation analysis of ABC Inc., which has been listed on the stock exchange for two years. The metrics to be used in your valuation include revenue, earnings per share (EPS), and dividends paid per share (DPS), which is presented in the table below:

Fiscal quarter		Fiscal year 1	Fiscal year 2
March	Revenue	\$3,784 (M)	\$4,097 (M)
	EPS	1.37	−0.34
	DPS	N/A	N/A
June	Revenue	\$4,236(M)	\$5,905(M)
	EPS	1.78	3.89
	DPS	N/A	0.25
September	Revenue	\$4,187(M)	\$4,997(M)
	EPS	−3.38	−2.88
	DPS	N/A	0.25
December	Revenue	\$3,889(M)	\$4,389(M)
	EPS	−8.66	−3.98
	DPS	N/A	0.25

The data available online are pre-organized into a tabular format.

Use the data to construct a two-dimensional rectangular array (i.e., data table) with the columns representing the metrics for valuation and the observations arranged in a time-series sequence.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.b] Describe how data are organized for quantitative analysis

Example: Organizing raw data into a data table

Answer:

Step 1: Determine the data table structure

- The columns should represent the three metrics: Revenue, EPS and DPS.
- The row should be the observations for each metrics in a time ordered sequence, starting from Q1 year 1 to Q4 year 2.

Step 2: Construct a two-dimensional array

	Revenue (\$ million)	EPS (\$)	DPS (\$)
Q1 Year 1	3,784	1.37	0
Q2 Year 1	4,236	1.78	0
Q3 Year 1	4,187	-3.38	0
Q4 Year 1	3,889	-8.66	0
Q1 Year 2	4,097	-0.34	0
Q2 Year 2	5,905	3.89	0.25
Q3 Year 2	4,997	-2.88	0.25
Q4 Year 2	4,389	-3.98	0.25

Now, the formatted data can be used in financial analysis and is readable by a computer.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.c] Interpret frequency and related distributions

1. Definition of Frequency distribution

A frequency distribution is a tabular display of data constructed either by counting the observations of a variable by distinct values or groups or by tallying the values of a numerical variable into a set of numerically ordered bins.

2. Constructing a frequency distribution

The following procedure describes how to construct a frequency distribution:

Example:

Use the data in Table to construct a frequency distribution for the returns on Intelco's common stock.

Table: Annual Returns for Intelco, Inc., Common Stock

10.4%	22.5%	11.1%	-12.4%
9.8%	17.0%	2.8%	8.4%
34.6%	-30%	0.6%	5.0%
-17.6%	5.6%	8.9%	40%
-1.0%	-4.2%	-5.2%	21.0%

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.c] Interpret frequency and related distributions

2. Constructing a frequency distribution

Step 1: Define the intervals

1.1. *Determine range of values:*

Sort the data in ascending order.



Calculate the range of the data, defined as
Range =
Maximum value –
Minimum value.

Example (cont.)

We determine the range of returns of Intelco's stock:

- Sort the data in ascending order:
–30%, –17.6%, –12.4%, –5.2%, –4.2%, –1.0%, 0.6%, 2.8%, 5.0%, 5.6%, 8.4%, 8.9%, 9.8%, 10.4%, 11.1%, 17.0%, 21.0%, 22.5%, 34.6%, 40%.
- The minimum observation is –30% and the maximum observation is +40%. So, the range is:
Range = $40\% - (-30\%) = 70\%$

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.c] Interpret frequency and related distributions

2.

Constructing a frequency distribution

Step 1: Define the intervals

1.2. *Determine the intervals:*

Decide on the number of bins (k) in the frequency distribution.

- If too few bins are used, the data may be too broadly summarized and important characteristics may be lost.
- If too many bins are used, the data may not be summarized enough.



Determine bin width as Range/k .

Example (cont.)

Decide on the number of bins and determine bin width:

- If we set $k = 70$, then the bin width is $70\%/70 = 1\%$, resulting in 70 separate bins, but it is too many data to summarize.
- Instead, we set $k = 7.0$, then the bin width is $70\%/7.0 = 10\%$.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.c] Interpret frequency and related distributions

2. Constructing a frequency distribution

Step 1: Define the intervals

1.2. Determine the intervals:

Determine the endpoints of the bins:

- Determine the first bin by adding the bin width to the minimum value.
- Then, determine the remaining bins by successively adding the bin width to the prior bin's end point and stopping after reaching a bin that includes the maximum value.

Example (cont.)

Determine the endpoints of the bins:

Beginning value		Bin width		Ending value
-30%	+	10%	=	-20%
-20%	+	10%	=	-10%
-10%	+	10%	=	0%
0%	+	10%	=	10%
10%	+	10%	=	20%
20%	+	10%	=	30%
30%	+	10%	=	40%

Note: The bins do not overlap, so each observation can be placed uniquely into one bin, and the last bin includes the maximum value.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.c] Interpret frequency and related distributions

2. Constructing a frequency distribution

Step 2: Tally and count the observations

Constructing a table with three columns:

We list the intervals in ascending order



We tally the information by assigning each observation to an interval.



We count and list the total of the number of observations, which is called **frequency**, that fall in each interval.



Interval

Tallies

Absolute Frequency

$-30\% \leq x < -20\%$

/

1

$-20\% \leq x < -10\%$

//

2

$-10\% \leq x < 0\%$

///

3

$0\% \leq x < 10\%$

//// /

7

$10\% \leq x < 20\%$

///

3

$20\% \leq x < 30\%$

//

2

$30\% \leq x \leq 40\%$

//

2

Total

20

→ The **model interval** is the interval with the greatest frequency. In this example, the model interval is $0\% \leq x < 10\%$, which includes 7 return observations.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.c] Interpret frequency and related distributions

2.

Constructing a frequency distribution

Step 2: Tally and count the observations

There are four types of frequency:

Absolute frequency

The actual number of observations counted for each unique value of the variable.



Relative frequency

The absolute frequency of each unique value of the variable divided by the total number of observations.



Cumulative absolute frequency

Created by adding up the absolute frequencies starting at the lowest interval and progressing through the highest.

Cumulative relative frequency

Created by adding up the relative frequencies starting at the lowest interval and progressing through the highest.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.c] Interpret frequency and related distributions

2.

Constructing a frequency distribution

Step 2: Tally and count the observations

Continuing with the example, we can also construct a table of other types of frequencies:

Interval	Absolute frequency	Cumulative absolute frequency	Relative frequency	Cumulative relative frequency
$-30\% \leq x < -20\%$	1	$0 + 1 = 1$	$1/20 = 0.05$ or 5%	$0\% + 5\% = 5\%$
$-20\% \leq x < -10\%$	2	$1 + 2 = 3$	$2/20 = 0.10$ or 10%	$5\% + 10\% = 15\%$
$-10\% \leq x < 0\%$	3	$3 + 3 = 6$	$3/20 = 0.15$ or 15%	$15\% + 15\% = 30\%$
$0\% \leq x < 10\%$	7	$6 + 7 = 13$	$7/20 = 0.35$ or 35%	$30\% + 35\% = 65\%$
$10\% \leq x < 20\%$	3	$13 + 3 = 16$	$3/20 = 0.15$ or 15%	$65\% + 15\% = 80\%$
$20\% \leq x < 30\%$	2	$16 + 2 = 18$	$2/20 = 0.10$ or 10%	$80\% + 10\% = 90\%$
$30\% \leq x \leq 40\%$	2	$18 + 2 = 20$	$2/20 = 0.10$ or 10%	$90\% + 10\% = 100\%$
Total	20		100%	

For example: For $x < 10\%$:

- The cumulative absolute frequency is 13.
- The cumulative relative frequency is 65%.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.d] Interpret a contingency table

1. Contingency table

A contingency table is a tabular format that displays the **frequency distributions of two or more categorical variables** simultaneously.

Example: Stock Frequencies by Sector and Market Capitalization

Sector Variable	Small	Mid	Large	Total
Communication service	55	35	20	110
Consumer Staples	50	30	30	110
Energy	175	95	20	290
Healthcare	275	105	55	435
Utilities	20	25	10	55
Total	575	290	135	1,000

Joint frequencies

Marginal frequencies
Sum of joint frequencies
across rows and columns

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.d] Interpret a contingency table

2.

Confusion matrix – an application of contingency table

- **Definition:** This is an analytical tool based on contingency table for evaluating the performance of a classification of model.
- **Composition:** A confusion matrix is in the form of an $n \times n$ (n features under consideration) table and the most common type of confusion matrix.

Example: Confusion Matrix for Bond Default Prediction Model

	Actual Yes	Actual No	Total
Predicted Yes	300	40	340
Predicted No	10	1,650	1,660
Total	310	1,690	2,000

Interpretation:

We can see that this classification model incorrectly predicts default in 40 cases where no default actually occurred and also incorrectly predicts no default in 10 cases where default actually did occur.

→ Probability of incorrectly predicts = $50/2,000 = 2.5\%$

→ Probability of correctly predicts = $(1,650 + 300)/2,000 = 97.5\%$

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.e] Describe ways that data may be visualized and uses of specific visualizations

Common types of data visualization

Histogram and frequency polygon

Bar chart

Tree map

Word cloud

Line chart

Scatter Plot

Heat Map

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

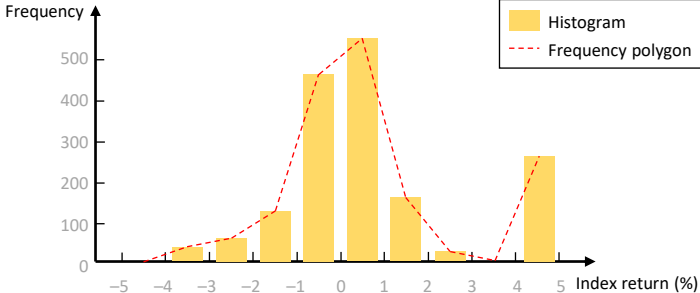
[LOS 2.e] Describe ways that data may be visualized and uses of specific visualizations

Histogram and frequency polygon

A **histogram** is a bar chart of data that have been grouped into a frequency distribution.

A **frequency polygon** is a graph of frequency distributions obtained by drawing straight lines joining successive midpoints of bars representing the class frequencies.

Example: Histogram Overlaid with Frequency Polygon for Daily Returns of EAA Equity Index



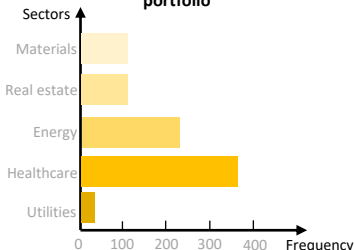
MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.e] Describe ways that data may be visualized and uses of specific visualizations

Bar chart

- A bar chart is used to plot the frequency distribution of categorical data, with each bar representing a distinct category and the bar's height (or length) proportional to the frequency of the corresponding category.
- Common types of bar chart: grouped bar chart, stacked bar chart, ...

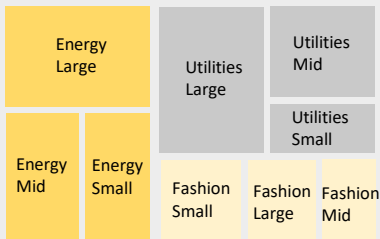
Example: Frequency by sector for stocks in a portfolio



Tree map

- A tree-map is a graphical tool to display categorical data which includes a set of colored rectangles to represent distinct groups.
- Additional dimensions of categorical data can be displayed by nested rectangles.
- The area of each rectangle is proportional to the value of the corresponding group.

Example: Tree-Map for Frequency Distribution by Sector in a Portfolio



MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.e] Describe ways that data may be visualized and uses of specific visualizations

Word cloud

A word cloud is a visual device for representing textual data, with the size of each distinct word being proportional to the frequency with which it appears in the given text.

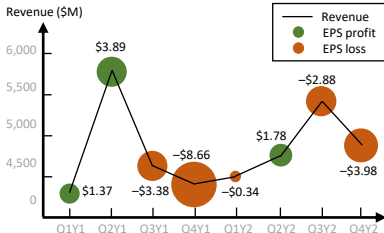
Example: Word Clouds Visualizing Excerpted Text in MDA Section in Form 10-Q of QXR Inc.



Line chart

- A line chart is a type of graph used to visualize ordered observations (often the change of data series over time).
- A *bubble line chart* is a special type of line chart that uses varying-sized bubbles as data points to represent an additional dimension of data.

Example: Quarterly revenue and EPS of ABC Incorporated



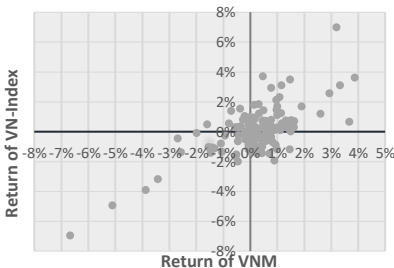
MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.e] Describe ways that data may be visualized and uses of specific visualizations

Scatter plot

A scatter plot is a type of graph for visualizing the joint variation in two numerical variables. It is a useful tool for displaying and understanding potential relationships between the variables.

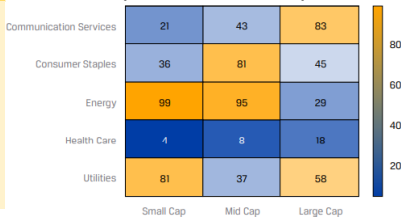
Example: Scatter Plot VMN stock return vs VN-Index return



Heat map

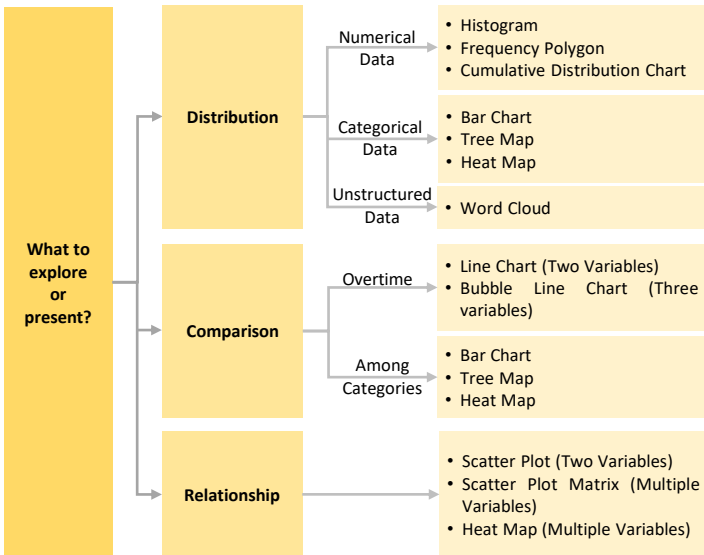
A heat map is a type of graphic that organizes and summarizes data in a tabular format and represents it using a color spectrum (often used in displaying frequency distributions or visualizing the degree of correlation among different variables).

Example: Frequencies by Sector and Market Capitalization in Heat Map



MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.f] Describe how to select among visualization types



MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.f] Describe how to select among visualization types

Example: A portfolio manager plans to buy several stocks traded on a small emerging market exchange but is concerned whether the market can provide sufficient liquidity to support her purchase order size. As the first step, she wants to analyze the daily trading volumes of one of these stocks over the past five years.

Explain which type of chart can best provide a quick view of trading volume for the given period.

Answer:

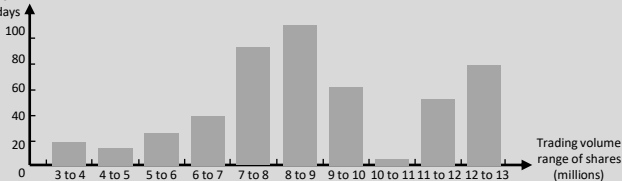
Step 1: Determine the purpose

The portfolio manager want to analyze the **daily trading volumes** of each stock that she plans to buy, so her purpose is to present distribution of data.

Step 2: Determine the types of data to present

The five-year history of daily trading volumes contains a large amount of numerical data. Therefore, a **histogram** is the best chart for grouping these data into frequency distribution bins and for showing a quick snapshot of the shape, center, and spread of the data's distribution.

Number of
trading days



MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.g] Calculate and interpret measures of central tendency

Basic concepts

- A **population** is defined as all members of a specified group.
- A **parameter** is any descriptive measure of a population.
- A **sample** is a subset of a population.
- A **sample statistic** is a quantity computed from or used to describe a sample.

Central tendency

Measures of central tendency identify the center, or average, of a data set. This central point can then be used to represent the typical, or expected, value in the data set which can be either a population or a sample.

Measures of central tendency

Median
(6)

Mean

Mode
(5)

Arithmetic mean
(1)

Weighted mean
(2)

Geometric mean
(3)

Harmonic mean
(4)

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.g] Calculate and interpret measures of central tendency

1.

Arithmetic mean

Definition: The arithmetic mean is the sum of the values of the observations divided by the number of observations.

Formula:

• Population mean:

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

(*N* is the number of observations in the population)

• Sample mean:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

(*n* is the number of observations in the sample)

Disadvantage of Arithmetic mean

Arithmetic mean is easily affected by outliers.

3 options for dealing with extreme values:

Option 1: Do nothing; use the data without any adjustment.

Option 2: Delete all the outliers to calculate **trimmed mean (1)**

Option 3: Replace the outliers with another value to calculate **winsorized mean (2)**

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.g] Calculate and interpret measures of central tendency

1.**Arithmetic mean****(1) Trimmed mean**

A trimmed mean exclude a stated small percentage of the lowest and highest values and then computing an arithmetic mean of the remaining values.

(2) Winsorized mean

A winsorized mean assign a stated percentage of the lowest values equal to one specified low value and a stated percentage of the highest values equal to one specified high value, and then it computes a mean from the restated data.

Or we can say:

A trimmed mean excludes the extreme observations.

Example: A 5% trimmed mean discards the lowest 2.5% and the highest 2.5% of values and computes the mean of the remaining 95% of values.

A winsorized mean substitute values for the extreme values.

Example: A 95% winsorized mean

- Setting the bottom 2.5% of values = the value at or below which 2.5% of all the values lie (2.5th percentile).
- Setting the top 2.5% of values = the value at or below which 97.5% of all the values lie (97.5th percentile).

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.g] Calculate and interpret measures of central tendency

1.

Arithmetic mean

Example: Trimmed mean and Winsorized mean

Consider Acme Corporation has the following returns of common stock: 2%, 10%, 12%, 13%, 15%, 17%, 17%, 18%, 19%, 42%.

With 20% trimmed mean and 80% winsorized mean, calculate the mean and number of observations.

Answer:

- Calculate Arithmetic mean:

$$\text{Arithmetic mean} = \frac{2\% + 10\% + 12\% + 13\% + 15\% + 17\% + 17\% + 18\% + 19\% + 42\%}{10} \\ = 16.5\%$$

We can see that arithmetic mean could be misleading as it is affected by outliers. Subsequently, in order to deal with extreme values, we use Trimmed mean or Winsorized mean.

- Calculate Trimmed mean:

Step 1: Determine the outliers

To trim the mean by 20%, we remove the lowest 10% and the highest 10% of values. The number lowest values is: $10 \times 10\% = 1$ and similarly, the number of highest value is 1. The result is that we eliminate the returns of 2% and 42%, leading to the number of observations are 8.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.g] Calculate and interpret measures of central tendency

1.

Arithmetic mean

Answer:

Step 2: Determine the trimmed mean

$$\text{Trimmed mean} = \frac{10\%+12\%+13\%+15\%+17\%+17\%+18\%+19\%}{8} = 15.125\%$$

• Winsorized mean:

Step 1: Determine the outliers

To winsorize the mean by 80%, we replace the lowest 10% and the highest 10% of values. In this example, we assume that we replace the smallest and largest values with their nearest observations.

2%	10%	12%	13%	15%	17%	17%	18%	19%	42%
↓									↓
10%	10%	12%	13%	15%	17%	17%	18%	19%	19%

Step 2: Determine the winsorized mean

$$\text{Winsorized mean} = \frac{10\%+10\%+12\%+13\%+15\%+17\%+17\%+18\%+19\%+19\%}{10} = 15\%$$

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.g] Calculate and interpret measures of central tendency

2.

Weighted mean

Definition: Weighted average is a calculation that takes into account the varying degrees of importance of the numbers in a data set.

Uses: It is frequently used to calculate the average return of a portfolio.

Formula: $\bar{X}_W = \sum_{i=1}^n w_i X_i$

Example: A portfolio consists of 50% common stocks, 40% bonds, and 10% cash. If the return on common stocks is 12%, the return on bonds is 7%, and the return on cash is 3%, what is the portfolio return?

Answer

$$\bar{X}_W = W_{\text{stock}}R_{\text{stock}} + W_{\text{bonds}}R_{\text{bonds}} + W_{\text{cash}}R_{\text{cash}}$$

$$\bar{X}_W = (0.50 \times 0.12) + (0.40 \times 0.07) + (0.10 \times 0.03) = 0.091 \text{ or } 9.1\%$$

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.g] Calculate and interpret measures of central tendency

3.

Geometric mean

Definition and uses: Geometric mean is frequently used to average rates of change over time or to calculate the growth rate of a variable over a period.

Formula:

- $\bar{X}_G = \sqrt[n]{X_1 * X_2 \dots X_n}$ with $X_i \geq 0$ for $i = 1, 2, \dots, n$.
- To calculate the geometric mean for investment returns data, we must add 1 to each return observation (expressed as a decimal) and then subtract 1 from the result:

$$1 + R_G = \sqrt[n]{(1 + R_1) * (1 + R_2) \dots (1 + R_n)}$$

Example: For the last three years, the returns for Acme Corporation common stock have been -9.34%, 23.45%, and 8.92%. Compute the compound annual rate of return over the 3-year period.

Answer:

$$\begin{aligned} 1 + R_G &= \sqrt[3]{(1 - 0.0934) * (1 + 0.2345) * (1 + 0.0892)} \\ &= (1.21903)^{1/3} = 1.06825 \\ \rightarrow R_G &= 6.825\% \end{aligned}$$

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.g] Calculate and interpret measures of central tendency

4.

Harmonic mean

Definition: The harmonic mean is a type of weighted mean in which an observation's weight is inversely proportional to its magnitude.

Uses:

- It is used as a measure of central tendency in the presence of outliers.
- It is used most often when the data consist of rates and ratios (i.e., P/Es).
- A well-known application of harmonic mean in the investment strategy is known as **cost averaging**.

Formula: $\bar{X}_H = \frac{n}{\sum_{i=1}^n \left(\frac{1}{X_i} \right)}$ with $X_i > 0$ for $i = 1, 2, \dots, n$.

Example: An investor purchases \$1,000 of mutual fund shares each month, and over the last three months the prices paid per share were \$8, \$9, and \$10. What is the average cost per share?

Answer:

$$\bar{X}_H = \frac{n}{\sum_{i=1}^n \left(\frac{1}{X_i} \right)} = \frac{3}{\frac{1}{8} + \frac{1}{9} + \frac{1}{10}} = 8.926$$

Note: Comparison between types of means:

Based on mathematical computation, we can see that for unequal values overtimes: Harmonic mean < Geometric mean < Arithmetic mean

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.g] Calculate and interpret measures of central tendency

5.**Mode**

- The mode is the value that occurs most frequently in a data set.
- A data set may have more than one mode or even no mode.

Example:

The mode of (1, 2, 4, 4, 6, 8, 10) is 4 because it occurs two times, which is more than any other number.

6.**Median**

- The median is the value of the middle item of a set of items that has been sorted into ascending or descending order.
- In an odd numbered sample of n items, the median is the value of the item that occupies the $(n + 1)/2$ position.
- In an even-numbered sample, we define the median as the mean of the values of items occupying the $n/2$ and $(n + 2)/2$ positions (the two middle items).

Example:

- The median of (1, 2, 4, 4, 6, 8, 10) is 4, which is the 4th position of dataset.
- The median of (1, 2, 4, 4, 6, 8, 10, 45) is 5, which is the mean value of the 4th and 5th position of dataset.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.h] Evaluate alternative definitions of mean to address an investment problem.

Deciding which central tendency measure to use

Collect sample

Include all values, including outliers?

Yes

Arithmetic mean

Compounding?

Yes

Geometric mean

Extreme outliers?

Yes

Harmonic mean
Trimmed mean
Winsorized mean

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.i] Calculate quantiles and interpret related visualization

Definition of quantile

Quantile is the general term for a value at or below which a stated proportion of the data in a distribution lies.

The most commonly used quantiles:

- Quartiles — the distribution is divided into quarters.
- Quintile — the distribution is divided into fifths.
- Decile — the distribution is divided into tenths.
- Percentile — the distribution is divided into hundredths (percents).

Formula

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

Where:

- y : the percentage point to divide the distribution.
- L_y : the location (L) of the percentile (P_y) in the array sorted in ascending order.
- n : the number of data points sorted in ascending order.

Note: As the sample size increases, the percentile location calculation becomes more accurate; in small samples it may be quite approximate.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

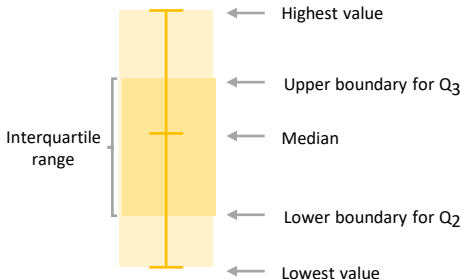
[LOS 2.i] Calculate quantiles and interpret related visualization

Box and whisker plot

Definition: Box and whisker plot is a type of diagram to visualize the dispersion of data across quartiles.

Composition: Box and whisker plot consists of a “box” with “whiskers” connected to the box.

- The box represents the lower bound of the second quartile and the upper bound of the third quartile.
- The whiskers are the lines that run from the box and are bounded by the “fences,” which represent the lowest and highest values of the distribution.



MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.i] Calculate quantiles and interpret related visualization

Quantiles in investment practice

Quantiles are used in portfolio performance evaluation as well as in investment strategy development and research

Rank performance

Example:

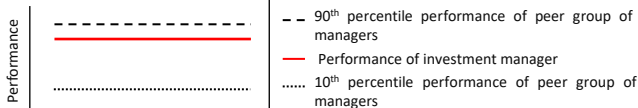
The performance of investment managers is often characterized in terms of the percentile or quartile in which they fall relative to the performance of their peer group of managers.

Investment research

Example:

Dividing data into quantiles based on some characteristic allows analysts to evaluate the impact of that characteristic on a quantity of interest.

Illustration of rank performance: Performance of investment managers



MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.j] Calculate and interpret measures of dispersion

Definition of dispersion

Dispersion is the variability around the central tendency. Or we can understand that if mean return addresses reward, then dispersion addresses risk.

Absolute dispersion

Absolute dispersion is the amount of variability present without comparison to any reference point or benchmark.

1. Range

2. Mean absolute deviation

3. Variance

4. Standard deviation

Relative dispersion

Relative dispersion is the amount of dispersion relative to a reference value or benchmark.

5. Coefficient of variation

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.j] Calculate and interpret measures of dispersion

1.

Range

Definition: The range is the difference between the maximum and minimum values in a dataset.

Formula: $\text{Range} = \text{Maximum value} - \text{Minimum value}$.

Application: In finance, range is used as a measure of volatility of a security. The size of the range corresponds to the security's level of risk: Safe securities, such as Government bonds, tend to show a smaller range of price.

Example:

What is the range for the 5-year annualized total returns for five investment managers if the managers' individual returns were 30%, 12%, 25%, 20%, and 23%?

Answer:

$\text{Range} = \text{Maximum value} - \text{Minimum value}$
 $= 30\% - 12\% = 18\%$

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.j] Calculate and interpret measures of dispersion

2.

Mean absolute deviation (MAD)

Definition: MAD is the average of the absolute values of the deviations of individual observations from the arithmetic mean.

Formula:
$$\text{MAD} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}$$

Application: MAD is a tool used to measure the forecasting error of a model.

Example: Find MAD of 5 following annualized returns achieved by an investment manager:

30%, 12%, 25%, 20%, 23% ? How is it interpreted?

Answer:

$$\bar{X} = \frac{30 + 12 + 25 + 20 + 23}{5} = 22\%$$

$$\text{MAD} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n} = \frac{|30-22| + |12-22| + |25-22| + |20-22| + |23-22|}{5} = 4.8\%$$

→ Interpretation: On average, an individual return will deviate $\pm 4.8\%$ from the mean return of 22%.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.j] Calculate and interpret measures of dispersion

3.

Variance

Definition: Variance is defined as the mean of the squared deviations from the arithmetic mean or from the expected value of a distribution.

Formula:

• Population variance: $\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$ • Sample variance: $s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$

4.

Standard deviation

Definition: Standard deviation is the positive square root of the variance and is frequently used as a quantitative measure of risk.

Formula:

• Population SD: $\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$ • Sample SD: $s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$

Application: Standard deviation is used as an indicator of market volatility and thus of risk.

Interpretation: The higher standard deviation, the greater the risk.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.j] Calculate and interpret measures of dispersion

Example: Assume that the 5-year annualized total returns for the five investment managers used in the preceding examples (30%, 12%, 25%, 20%, 23%) represent only a sample of the managers at a large investment firm. What is the sample variance and standard deviation of these returns?

Answer:

- Calculate sample mean:

$$\bar{X} = \frac{30 + 12 + 25 + 20 + 23}{5} = 22\%$$

- Calculate sample variance and sample standard deviation:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$
$$= \frac{(30 - 22)^2 + (12 - 22)^2 + (25 - 22)^2 + (20 - 22)^2 + (23 - 22)^2}{5 - 1} = 44.5 (\%^2)$$

$$\rightarrow s = \sqrt{44.5 \%^2} = 6.67\%$$

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.j] Calculate and interpret measures of dispersion

5.

Coefficient of variation (CV)

Definition: The coefficient of variation (CV) is the ratio of the standard deviation of a set of observations to their mean value.

Use: By expressing the magnitude of variation among observations relative to their average size, the CV permits direct comparisons of dispersion across different datasets.

Formula: $CV = \frac{s}{\bar{X}}$

Interpretation: The larger CV, the higher risk (riskier).

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.j] Calculate and interpret measures of dispersion

5.

Coefficient of variation (CV)

Example: The mean monthly return and the standard deviation for three industry sectors are shown in the following exhibit.

Sector	Mean monthly return (%)	Standard deviation of return (%)
Utilities	2.10	1.23
Consumer goods	1.25	1.35
Communication	3.01	1.52

Based on the coefficient of variation, which is the most attractive sector?

Answer:

$$CV_{\text{Utilities}} = \frac{1.23\%}{2.10\%} = 0.59; \quad CV_{\text{Consumer goods}} = \frac{1.35\%}{1.25\%} = 1.08;$$

$$CV_{\text{Communication}} = \frac{1.52\%}{3.01\%} = 0.51$$

→ Communication sector has the lowest CV so it is the most attractive sector.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.k] Calculate and interpret target downside deviation

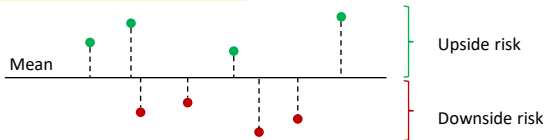
Variance and Standard deviation take account of:

Upside risk

Risk measure based on outcome *above* the mean.

Downside risk

Risk measure based on outcome *below* the mean.



Target downside deviation (target semideviation)

Definition: The target downside deviation, also referred to as the target semideviation, is a measure of downside risk.

Formula:

$$s_{\text{Target}} = \sqrt{\sum_{\text{for all } X_i \leq B}^n \frac{(X_i - B)^2}{n - 1}}$$

where: B is the target under consideration, or we can say it is the minimum target return.

Interpretation: The larger target semideviation, the higher risk.

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.k] Calculate and interpret target downside deviation

Example:

A fund had the following experience over the past 10 years:

Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
4.5%	6.0%	1.5%	-2%	0%	4.5%	3.5%	2.5%	5.5%	4.0%

Find the target semideviation of the returns over 10 years if:

Case 1: The target is 2%?

Case 2: The target is 3%?

Answer:

Case 1: $X_i \leq B = X_i \leq 2\% = \{-2\%; 0.0\%; 1.5\%\}$

$$s_{\text{Target}} = \sqrt{\sum_{\text{for all } X_i \leq B} \frac{(X_i - B)^2}{n-1}}$$

$$= \sqrt{\frac{(-2.0\% - 2\%)^2 + (0.0\% - 2\%)^2 + (1.5\% - 2\%)^2}{9}} = 1.5\%$$

Case 2: $X_i \leq B = X_i \leq 3\% = \{-2\%; 0.0\%; 1.5\%; 2.5\%\}$

$$s_{\text{Target}} = \sqrt{\frac{(-2.0\% - 3\%)^2 + (0.0\% - 3\%)^2 + (1.5\% - 3\%)^2 + (2.5\% - 3\%)^2}{9}} \approx 2.01\%$$

→ The higher target, the larger target semideviation, and as a result the higher risk that the target could not be met.

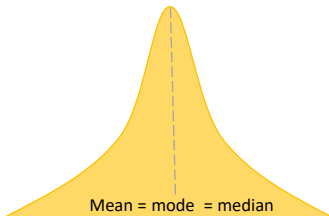
MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.I] Interpret skewness

Symmetrical distribution

Definition: A distribution is symmetrical if it is shaped identically on both sides of its mean. Distributional symmetry implies that intervals of losses and gains will exhibit the same frequency.

Example: A symmetrical distribution with a mean return of zero will have losses in the -6% to -4% interval as frequently as it will have gains in the $+4\%$ to $+6\%$ interval.



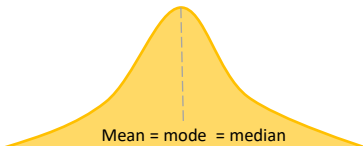
More information than the mean and variance is needed to characterize its shape.

Normal distribution

Definition: Normal distribution is a probability distribution that is symmetric about the mean (bell-shaped), showing that data near the mean are more frequent in occurrence than data far from the mean.

Features of normal distribution:

- Its mean, median, and mode are equal.
- It is completely described by two parameters—its mean and variance.
- Skewness = 0 & kurtosis = 3.



Completely described by two parameters—its mean and variance

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.I] Interpret skewness

Definition of skewness

Definition: Skewness refers to the extent to which a distribution is not symmetrical. Nonsymmetrical distributions may be either positively or negatively skewed and result from the occurrence of outliers in the data set.

Formula: Sample skewness = $\frac{1}{n} \times \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$

Note: The skewness of symmetrical distributions is equal to 0.

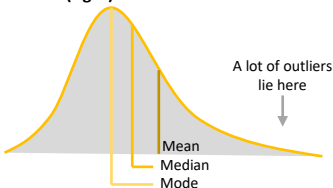
Positively skewed (skewness > 0)

- A positively skewed distribution is characterized by many outliers in the upper region, or right tail.
- A positively skewed distribution is said to be skewed right because of its relatively *long upper (right) tail*.

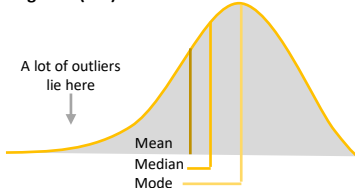
Negatively skewed (skewness < 0)

- A negatively skewed distribution is characterized by many outliers that fall within its lower (left) tail.
- A negatively skewed distribution is said to be skewed left because of its *long lower (left) tail*.

Positive (right) skew: mode < median < mean



Negative (left) skew: mode > median > mean



MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.m] Interpret kurtosis

Kurtosis

Definition:

- Kurtosis is a measure of the degree to which a distribution is more or less “peaked” than a normal distribution.
- Excess kurtosis** is the kurtosis relative to the normal distribution.

Formula:

$$\text{Sample kurtosis} = \frac{1}{n} \times \frac{\sum_{i=1}^N (X_i - \bar{X})^4}{s^4}$$

$$\rightarrow \text{Excess kurtosis} = \text{Kurtosis} - 3$$

Note: The Kurtosis of a normal distribution is equal to 3.

Degree of kurtosis

Platykurtic (kurtosis < 3)

This term describes a distribution that is less peaked, or *thinner* than a normal distribution.

$$\rightarrow \text{excess kurtosis} < 0$$

Mesokurtic (kurtosis = 3)

This term describes a distribution that has the *same* kurtosis as a normal distribution.

$$\rightarrow \text{excess kurtosis} = 0$$

Leptokurtic (kurtosis > 3)

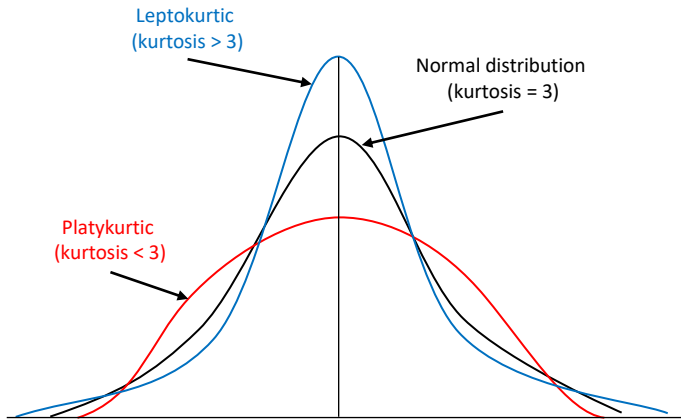
This term describes a distribution that is more peaked, or *fatter* than a normal distribution.

$$\rightarrow \text{excess kurtosis} > 0$$

MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.m] Interpret kurtosis

Degree of kurtosis



MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.n] Interpret correlation between two variables

Basics concept of correlation coefficient

Definition: The correlation coefficient is a statistic that measures the association between two variables.

Interpretation:

- A positive correlation coefficient indicates that the two variables tend to move together.
- A negative correlation coefficient indicates that the two variables tend to move in opposite directions.

Formula

Sample correlation coefficient:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \quad \text{Where: } s_{XY} (\text{sample covariance}) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1} \quad (*)$$

Interpretation

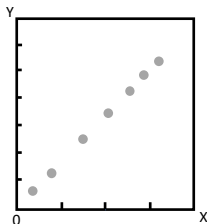
Correlation coefficient range from -1 to 1

- $r_{XY} = 0 \Rightarrow$ No linear relation ship between X and Y .
- $r_{XY} = 1 \Rightarrow$ Perfect positive linear relationship between X and Y .
- $r_{XY} = -1 \Rightarrow$ Perfect negative linear relationship between X and Y .

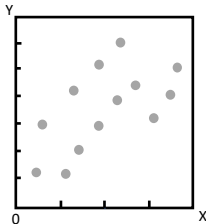
MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.n] Interpret correlation between two variables

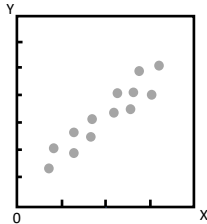
Scatterplot showing various degrees of correlation



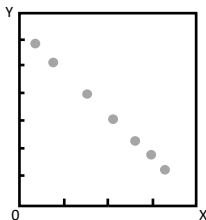
Perfect positive correlation



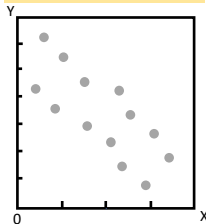
Low degree of positive correlation



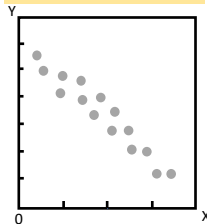
High degree of positive correlation



Perfect negative correlation



Low degree of negative correlation

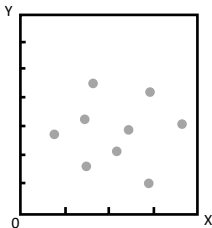


High degree of negative correlation

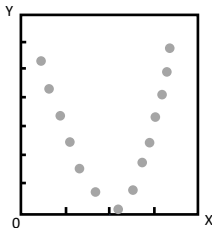
MODULE 2: ORGANIZING, VISUALIZING AND DESCRIBING DATA

[LOS 2.n] Interpret correlation between two variables

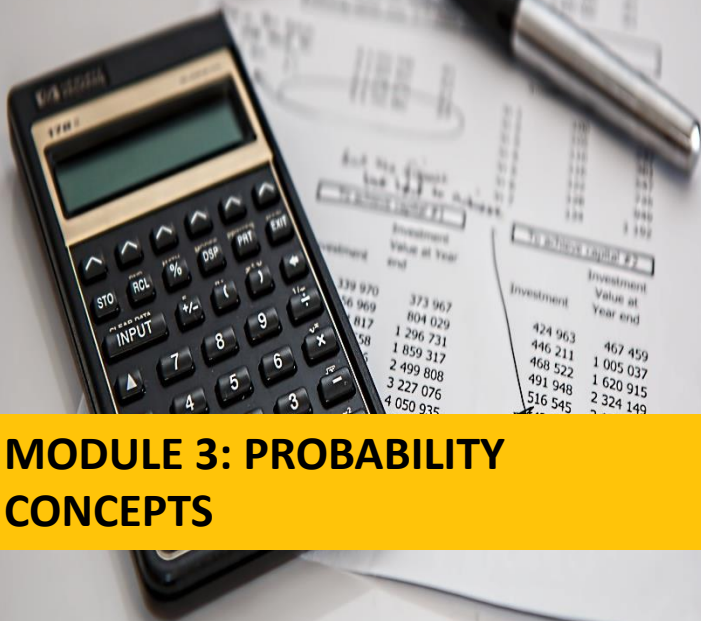
Scatterplot showing various degrees of correlation



No correlation



Non-linear correlation



MODULE 3: PROBABILITY CONCEPTS

MODULE 3: PROBABILITY CONCEPTS

Learning outcome statements

[LOS 3.a] Define a random variable, an outcome, and an event.

[LOS 3.b] Identify the two defining properties of probability, including mutually exclusive and exhaustive events, and compare and contrast empirical, subjective, and a priori probabilities.

[LOS 3.c] Describe the probability of an event in terms of odds for and against the event.

[LOS 3.d] Calculate and interpret conditional probabilities

[LOS3.e] Demonstrate the application of the multiplication and addition rules for probability

[LOS 3.f] Compare and contrast dependent and independent events.

[LOS 3.g] Calculate and interpret an unconditional probability using the total probability rule.

[LOS 3.h] Calculate and interpret the expected value, variance, and standard deviation of random variables.

MODULE 3: PROBABILITY CONCEPTS

Learning outcome statements

[LOS 3.i] Explain the use of conditional expectation in investment applications

[LOS 3.j] Interpret a probability tree and demonstrate its application to investment problems.

[LOS 3.k] Calculate and interpret the expected value, variance, standard deviation, covariances, and correlations of portfolio returns

[LOS 3.l] Calculate and interpret the covariances of portfolio returns using the joint probability function.

[LOS 3.m] Calculate and interpret an updated probability using Bayes' formula.

[LOS 3.n] Identify the most appropriate method to solve a particular counting problem and analyze counting problems using factorial, combination, and permutation concepts.

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.a] Define a random variable, an outcome and an event

Basic concepts

Random variable

A quantity whose future outcomes are uncertain.

Outcome

A possible value of a random variable.

Event

An outcome or a specified set of outcomes.

Probability

Likelihood of an event happening.

Example

Consider rolling a 6-sided die one time.

- Random variable: The numbers that come up by rolling the dice.
- Outcome: Randomly getting a 4 in a roll.
- Event: Getting 4 more than two times in six rolls.
- Probability: The probability of rolling any one of the numbers 1 to 6 with a fair die is $1/6 = 0.1667$ or 16.7%.

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.b] Identifying mutually exclusive and exhaustive events, and compare and contrast empirical, subjective , and a priori probability

1.

Basic features of probability

- The probability of any event E is a number between 0 and 1: $0 \leq P(E) \leq 1$.
- The sum of the probabilities of any set of **mutually exclusive and exhaustive events** equals 1.

2.

Mutually exclusive vs exhaustive probability

Mutually exclusive

Definition: Mutually exclusive events are events that cannot occur at the same time.

Visualization: A & B is mutually exclusive when $P(A \cap B) = 0$



Exhaustive

Definition: Exhaustive means that the events cover **all** possible outcomes.

Visualization:



All possible outcomes

Note: A & B can overlap each other

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.b] Identifying mutually exclusive and exhaustive events, and compare and contrast empirical, subjective , and a priori probability

3.

Types of probability

Types of
probability

Subjective probability (Personal judgement)

Objective
probability**Empirical probability**

An empirical probability is established by analyzing past data.

Priori probability

An a priori probability is determined using a formal reasoning and inspection process.

Example:

- **Subjective probability:** Statements such as “I believe there is a 70% probability that Acme Foods will outperform the market this year” is a subjective probability.
- **Empirical probability:** For a stock, based on prior patterns of up and down days, the probability of the stock having a down day tomorrow is an empirical probability.
- **Priori probability:** On a random draw, the probability of choosing a stock of a particular industry from the S&P 500 is a priori probability.

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.c] Describe the probability of an event in terms of odds for and against the event

Odds for

- The **odds for** an event are the ratio of the number of ways the event can occur to the number of ways the event can not occur.

$$\text{Odds for } E = \frac{P(E)}{1 - P(E)}$$

where $P(E)$ is the probability of event.

- Given odds of a to b :

$$\text{Odds for } E = \frac{a}{a + b}$$

Odds against

- The **odds against** an event are the ratio of the number of ways the event can not occur to the number of ways the event can occur.

$$\text{Odds against } E = \frac{1 - P(E)}{P(E)}$$

where $P(E)$ is the probability of event

- Given odds against of a to b :

$$\text{Odds against } E = \frac{b}{a + b}$$

Example:

Considering an event that has a probability of occurrence of 0.125. Calculate the odds for this event and the odds against this event.

Answer:

$$\text{The odds for this event} = \frac{P(E)}{1 - P(E)} = \frac{0.125}{1 - 0.125} = \frac{1}{7} \text{ (Stated: one to seven)}$$

$$\text{The odds against this event} = \frac{1 - P(E)}{P(E)} = \frac{1 - 0.125}{0.125} = 7 \text{ (Stated: seven to one)}$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.d] Calculate and interpret conditional probabilities

Unconditional probability

Definition: An unconditional probability is the probability of an event not conditioned on another event.

Denotation: $P(B)$ – also known as marginal probability of B

Conditional probability

Definition: A conditional probability is the probability of an event given another event.

Denotation: $P(A|B)$ – Probability of an event A given event B. The occurrence of B affects the occurrence of A, so we have the following formula:

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

where $P(AB)$ is the **joint probability** that A & B happen together.

Example:

If event B happens half the time, $P(B) = 0.5$, and event A and B both happen 10% of the time, $P(AB) = 0.1$. What is the probability that A happens, given B happens?

Answer:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{0.1}{0.5} = 0.2 \text{ or } 20\%$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.e] Demonstrate the application of the multiplication rule and addition rule for probabilities & [LOS 3.f] Compare and contrast dependent and independent events

1.

Basic definitions of dependent & independent events

- When events are **independent**, the occurrence of one event does not affect the probability of occurrence of the other event.

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B)$$

- Otherwise, the events are **dependent**.

2.

Rules of probability calculation

Addition rule

- Not mutually exclusive events:
 $P(A \text{ or } B) = P(A) + P(B) - P(AB)$
- Mutually exclusive events:
 $P(A \text{ or } B) = P(A) + P(B)$

Multiplication rule

- Dependent events:
 $P(AB) = P(A|B)P(B) = P(B|A)P(A)$
- Independent events:
 $P(AB) = P(A)P(B)$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.e] Demonstrate the application of the multiplication rule and addition rule for probabilities & [LOS 3.f] Compare and contrast dependent and independent events

3.

Example of addition rule

For events that are not mutually exclusive

Example: Consider the following information:

- $P(I) = 0.4$, the probability of the monetary authority increasing interest rates (I) is 40%.
- $P(RI) = 0.28$, the probability of a recession and an increase in interest rates happening at the same time is 28%.
- The probability of a recession, $P(R)$, is 34%.

Determine the probability that either interest rates will increase or a recession will occur.

Answer:

R and I is not mutually exclusive because they can happen at the same time.

$$\begin{aligned}\rightarrow P(R \text{ or } I) &= P(R) + P(I) - P(RI) \\ &= 0.34 + 0.40 - 0.28 = 0.46\end{aligned}$$

For mutually exclusive events

Example: A single 6-sided die is rolled. What is the probability of rolling a 2 or a 5?

Answer:

In this case the probability of rolling a 2 and a 5 are mutually exclusive events because there is no way that they happen together in a single roll.

$$\rightarrow P(2 \text{ or } 5) = P(2) + P(5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.e] Demonstrate the application of the multiplication rule and addition rule for probabilities & [LOS 3.f] Compare and contrast dependent and independent events

4.

Example of multiplication rule

For dependent events

Example: Helen Pedersen has all her money invested in either of two mutual funds (A and B). She knows that there is a 40% probability that fund A will rise in price and a 60% chance that fund B will rise in price if fund A rises in price. What is the probability that both fund A and fund B will rise in price?

Answer:

The increase in price of fund A is related to the increase in price of fund B

→ These are 2 dependent events

→ $P(A \text{ and } B \text{ both increase}) = P(A \text{ increase}) \times P(B \text{ increase given } A \text{ increase})$
 $= 0.40(0.60) = 0.24$

For independent events

Example: What is the probability of rolling three 4s in one simultaneous toss of three dice?

Answer:

- The result of the previous rolling does not dictate the result of the following rolling → The result of rolling dices is independent events.
- Since the probability of rolling a 4 for each die is $\frac{1}{6}$ and each roll is independent of each other, the probability of rolling three 4s is:

$$P(\text{three times } 4) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{216} = 0.00463$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.g] Calculate and interpret an unconditional probability using total probability rule

The total probability rule

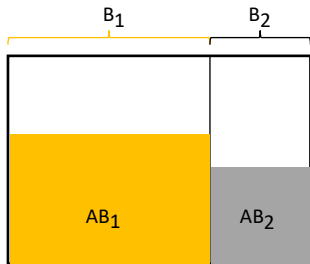
The total probability rule is used to determine the unconditional probability of an event, given conditional probabilities.

Case 1: Total probability rule for two scenarios

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$$

Where:

- $P(B_1) + P(B_2) = 1$
- B_1 and B_2 is a mutually exclusive and exhaustive set of outcomes.

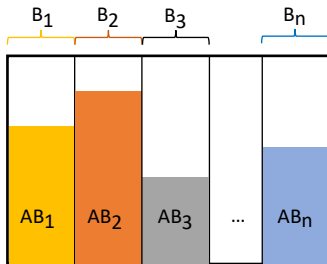


Case 2: Total probability rule for n scenarios

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_N)P(B_N)$$

Where:

- $\sum_{i=1}^n P(B_i) = 1$
- B_1, B_2, \dots, B_N is a mutually exclusive and exhaustive set of outcomes.



MODULE 3: PROBABILITY CONCEPTS

[LOS 3.g] Calculate and interpret an unconditional probability using total probability rule

Example: You have developed a set of criteria for evaluating distressed credits. Companies that do not receive a passing score are classed as likely to go bankrupt within 12 months. You gathered the following information when validating the criteria:

- Forty percent of the companies to which the test is administered will go bankrupt within 12 months: $P(\text{nonsurvivor}) = 0.40$.
- Fifty-five percent of the companies to which the test is administered pass it: $P(\text{pass test}) = 0.55$.
- The probability that a company will pass the test given that it will subsequently survive 12 months, is 0.85: $P(\text{pass test} | \text{survivor}) = 0.85$. What is $P(\text{pass test} | \text{nonsurvivor})$?

Answer:

Step 1: Determine components of total probability rule

- A is supposed to be “pass test” while B_1 is “nonsurvivor” and B_2 is “survivor”.
- Because B_1 and B_2 is mutually exclusive and exhaustive
 $\rightarrow P(B_1) + P(B_2) = 1 \rightarrow P(B_2) = 0.6$

Step 2: Apply total probability rule

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$$

$$\rightarrow P(\text{pass test}) =$$

$$P(\text{pass test} | \text{nonsurvivor})P(\text{nonsurvivor}) + P(\text{pass test} | \text{survivor})P(\text{survivor}) \\ = P(\text{pass test} | \text{nonsurvivor})(0.40) + 0.85(0.60) = 0.55$$

$$\rightarrow P(\text{pass test} | \text{nonsurvivor}) = \frac{0.55 - (0.85 \times 0.6)}{0.40} = 0.10 \text{ or } 10\%.$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.h] Calculate and interpret the expected value, variance, and standard deviation of random variables

1. Expected value of a random variable

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

Usage: Expected value (for example, expected stock return) looks either to the future, as a forecast, or to the “true” value of the mean (the population mean).

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

Example:

Suppose a random variable can take one of these values $X = \{5, 10\}$ with $P(5) = 40\%$ and $P(10) = 60\%$. Calculate the expected value $E(X)$.

Answer:

$E(X) = P(X_1)X_1 + P(X_2)X_2 = 40\% \times 5 + 60\% \times 10 = 8.$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.h] Calculate and interpret the expected value, variance, and standard deviation of random variables

2.

Variance and standard deviation of a random variable

Definition:

- The variance of a random variable is the expected value (the probability-weighted average) of squared deviations from the random variable's expected value.
- Standard deviation is the positive square root of variance.

Formula:

$$\begin{aligned}\sigma^2(X) &= P(X_1)[X_1 - E(X)]^2 + P(X_2)[X_2 - E(X)]^2 + \dots + P(X_n)[X_n - E(X)]^2 \\ &= \sum_{i=1}^n P(X_i) [X_i - E(X)]^2 \\ \sigma(X) &= \sqrt{\sigma^2(X)}\end{aligned}$$

Example:

Continue with previous example, we have outcomes of $X = \{5, 10\}$ with $P(5) = 40\%$ and $P(10) = 60\%$. Calculate the variance $\sigma^2(X)$ and standard deviation $\sigma(X)$.

Answer:

$$\begin{aligned}\sigma^2(X) &= P(X_1)[X_1 - E(X)]^2 + P(X_2)[X_2 - E(X)]^2 \\ &= 40\% \times (5 - 8)^2 + 60\% \times (10 - 8)^2 = 6. \\ \sigma(X) &= \sqrt{\sigma^2(X)} = \sqrt{6} \approx 2.45.\end{aligned}$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.i] Explain the use of conditional expectation in investment decisions

When we refine our expectations or forecasts, we are typically making adjustments based on new information or events.



The use of **conditional expected values** for each outcome is necessary.



The goal is to calculate the expected value of a random variable given all the possible outcomes that can occur.



Expected value for future performance can be calculated using total probability rule:

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

where S_1, S_2, \dots, S_n are mutually exclusive and exhaustive outcomes.

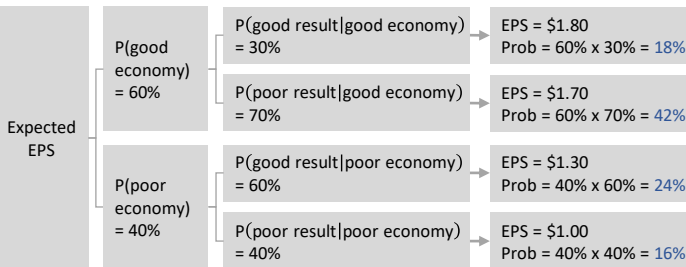
MODULE 3: PROBABILITY CONCEPTS

[LOS3.j] Interpret a probability tree and demonstrate its application

Basic function of a probability tree:

- Tree diagrams are a way of showing combinations of two or more events.
- Each branch is labelled at the end with its outcome and the probability is written alongside the line.

Example: Scenarios for a company's EPS using a probability tree



Interpretation:

$$\begin{aligned}
 \text{Expected EPS} &= \sum \text{Expected result of each scenario} \times \text{probability of each scenario} \\
 &= 1.8 \times 18\% + 1.70 \times 42\% + 1.30 \times 24\% + 1 \times 16\% = \$1.51
 \end{aligned}$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.k] Calculate and interpret the expected value, variance, standard deviation, covariance, and correlation of portfolio returns

1.

Expected return of a portfolio returns

Interpretation: Expected return is a measure used to evaluate potential profitability of investment opportunities.

Formula: $E(R_p) = E(w_1R_1 + w_2R_2 + \dots + w_nR_n)$
 $= w_1E(R_1) + w_2E(R_2) + \dots + w_nE(R_n)$

where w_i , R_i ($1 \leq i \leq n$) represent weight and return of individual security in the portfolio.

Example:

Suppose we have estimated expected returns on assets in the three-asset portfolio shown in table below. Calculate expected return of the portfolio.

Asset class	Weight	Expected return (%)
S&P 500	0.5	13
US long-term corporate bonds	0.25	6
MSCI EAFE	0.25	15

Answer:

The expected return of the portfolio is calculated as follow:

$$E(R_p) = w_1E(R_1) + w_2E(R_2) + \dots + w_nE(R_n)$$
$$= 0.5 \times 13\% + 0.25 \times 6\% + 0.25 \times 15\% = 11.75\%$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.k] Calculate and interpret the expected value, variance, standard deviation, covariance, and correlation of portfolio returns

2.

Covariance on a portfolio returns

Definition: Covariance is a measure of how two assets move together. It is the expected value of the product of the deviations of the two random variables from their respective expected values.

Basic characteristics:

- The covariance of R_i with itself is equal to variance of R_i :

$$\text{Cov}(R_i, R_i) = \text{Var}(R_i) = \sigma_{R_i}^2$$

- The covariance may range from negative infinity to positive infinity.

Formula: Covariance from a probability model

$$\text{Cov}(R_i, R_j) = E \left[(R_i - ER_i)(R_j - ER_j) \right]$$

where:

- R_i and R_j are two random variables.
- ER_i and ER_j are expected return of a portfolio returns.

MODULE 3: PROBABILITY CONCEPTS

[LOS3.I] Calculate and interpret the covariance of portfolio returns using the joint probability function

Example: The joint probabilities of the returns of Asset A and Asset B are given in the following figure. Calculate the covariance of returns for Asset A and Asset B.

Joint probability	Scenario 1: $R_B = 0.40$	Scenario 2: $R_B = 0.20$	Scenario 3: $R_B = 0$
Scenario 1: $R_A = 0.20$	0.15	0	0
Scenario 2: $R_A = 0.15$	0	0.60	0
Scenario 3: $R_A = 0.04$	0	0	0.25

Answer:

A and B happen together $\rightarrow P(A_n) = P(A_n, B_n)$

Step 1: Calculate expected return for the individual asset

$$E(R_A) = P(\text{Scenario}_1)R_A + P(\text{Scenario}_2)R_A + P(\text{Scenario}_3)R_A \\ = (0.15)(0.20) + (0.60)(0.15) + (0.25)(0.04) = 0.13 \text{ or } 13\%$$

$$E(R_B) = P(\text{Scenario}_1)R_B + P(\text{Scenario}_2)R_B + P(\text{Scenario}_3)R_B \\ = (0.15)(0.40) + (0.60)(0.20) + (0.25)(0.00) = 0.18 \text{ or } 18\%$$

Step 2: Calculate the covariance of assets returns

Joint probability	R_A	R_B	$P \times [R_A - E(R_A)] \times [R_B - E(R_B)]$
0.15	0.20	0.40	$0.15 \times (0.20 - 0.13) \times (0.40 - 0.18) = 0.00231$
0.60	0.15	0.20	$0.60 \times (0.15 - 0.13) \times (0.20 - 0.18) = 0.00024$
0.25	0.04	0.00	$0.25 \times (0.04 - 0.13) \times (0.00 - 0.18) = 0.00405$

\rightarrow Covariance of returns for Asset A and Asset B is:

$$0.00231 + 0.00024 + 0.00405 = 0.0066$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.k] Calculate and interpret the expected value, variance, standard deviation, covariance, and correlation of portfolio returns

3.

Correlation of a portfolio returns (*)

Formula:

$$\text{Corr}(R_i, R_j) = \rho(R_i, R_j) = \frac{\text{Cov}(R_i, R_j)}{\sigma(R_i)\sigma(R_j)}$$

$$\text{Corr}(R_i, R_j) \in (-1, 1)$$

Example:

Continue with previous example, we have:

Joint probability	$R_{B1} = 0.40$	$R_{B2} = 0.20$	$R_{B3} = 0$
$R_{A1} = 0.20$	0.15	0	0
$R_{A2} = 0.15$	0	0.60	0
$R_{A3} = 0.04$	0	0	0.25

- $\text{Cov}(R_A, R_B) = 0.0066$.
- $E(R_A) = 0.13$ and $E(R_B) = 0.18$.

Calculate correlation of the portfolio returns.

() Definition of Correlation has been mentioned in LOS 2.n, Module 2*

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.k] Calculate and interpret the expected value, variance, standard deviation, covariance, and correlation of portfolio returns

3.

Correlation of a portfolio returns (*)

Answer:

Step 1: Calculate standard deviation of the returns of Asset A and Asset B

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

=

$$\sqrt{P(R_{A1}, R_{B1})[R_{A1} - E(R_A)]^2 + P(R_{A2}, R_{B2})[R_{A2} - E(R_A)]^2 + P(R_{A3}, R_{B3})[R_{A3} - E(R_A)]^2}$$

$$= \sqrt{0.15 \times (0.20 - 0.13)^2 + 0.60 \times (0.15 - 0.13)^2 + 0.25 \times (0.04 - 0.13)^2} = 0.055$$

$$\bullet \sigma(R_B) = \sqrt{\sigma^2(R_B)}$$

$$= \sqrt{P(R_{A1}, R_{B1})[R_{B1} - E(R_B)]^2 + P(R_{A2}, R_{B2})[R_{B2} - E(R_B)]^2 + P(R_{A3}, R_{B3})[R_{B3} - E(R_B)]^2}$$

$$= \sqrt{0.15 \times (0.40 - 0.18)^2 + 0.60 \times (0.20 - 0.18)^2 + 0.25 \times (0 - 0.18)^2} = 0.125$$

Step 2: Calculate correlation

$$\rho(R_A, R_B) = \frac{\text{Cov}(R_A, R_B)}{\sigma(R_A)\sigma(R_B)} = \frac{0.0066}{0.055 \times 0.125} = 0.96$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.k] Calculate and interpret the expected value, variance, standard deviation, covariance, and correlation of portfolio returns

4.

Variance of a portfolio returns

Formula:

$$\text{Var}(R_p) = \sigma_p^2 = \sum_{i=1}^N \sum_{j=1}^N w_i w_j \text{Cov}(R_i, R_j)$$

→ The variance of a portfolio composed of risky asset A and risky asset B:

$$\begin{aligned}\sigma_p^2 &= W_A^2 \sigma^2(R_A) + W_B^2 \sigma^2(R_B) + 2W_A W_B \text{Cov}(R_A, R_B) \\ &= W_A^2 \sigma^2(R_A) + W_B^2 \sigma^2(R_B) + 2W_A W_B \rho(R_A, R_B) \sigma(R_A) \sigma(R_B)\end{aligned}$$

Example:

A portfolio is 30% invested in stocks with a standard deviation of returns of 20%, and the remainder is invested in bonds with a standard deviation of returns of 12%. The correlation of bond returns with stock returns is 0.6. Calculate the standard deviation of returns for the portfolio.

Answer:

Portfolio standard deviation is:

$$\begin{aligned}\sigma_p &= \sqrt{w_{\text{stock}}^2 \sigma_{\text{stocks}}^2 + w_{\text{bonds}}^2 \sigma_{\text{bonds}}^2 + 2w_{\text{stocks}} w_{\text{bonds}} \sigma_{\text{stocks}} \sigma_{\text{bonds}} \rho(\text{stocks}, \text{bonds})} \\ &= \sqrt{(0.3^2)(0.20^2) + (0.7^2)(0.12^2) + 2(0.3)(0.7)(0.20)(0.12)(0.60)} \\ &= \sqrt{0.0167} = 12.92\%\end{aligned}$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.m] Calculate and interpret an updated probability using Bayes' formula

1.

Why we need to use Bayes' formula

When we make decisions involving investments, the viewpoints, which we start based on experience and knowledge, might be changed or confirm by new observations.

→ **Bayes' formula** is a rational method for adjusting our viewpoints as we confront new information.

2.

General rule

Updated probability = $\frac{\text{probability of new information for a given event}}{\text{unconditional probability of new information}}$ x prior probability event,

$$\text{or: } P(\text{Event}|\text{Information}) = \frac{P(\text{Information}|\text{Event})}{P(\text{Information})} \times P(\text{Event})$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.m] Calculate and interpret an updated probability using Bayes' formula

3.

Deriving the rule

Begin with multiplication rule of dependent events:

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

$$P(BA) = P(B|A) \times P(A)$$

Since the joint probabilities $P(AB)$ and $P(BA)$ are equal, we can equate expressions $P(A|B) \times P(B) = P(B|A) \times P(A)$

Replacing B with Event, and A with Information, Bayes' formula can be translated to:

$$P(\text{Event}|\text{Information}) = \frac{P(\text{Information}|\text{Event})}{P(\text{Information})} \times P(\text{Event})$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.m] Calculate and interpret an updated probability using Bayes' formula

Example:

There is a 30% probability the economy will outperform, and if it does, there is a 70% chance a stock will go up and a 30% chance the stock will go down.

There is a 40% probability the economy will meet expectation, and if it does, there is a 50% chance a stock will go up and a 50% chance the stock will go down.

There is a 30% chance the economy will underperform, and if it does, there is a 20% chance the stock in question will increase in value (have gains) and an 80% chance it will not.

Given that the stock increased in value, calculate the probability that the economy outperformed.

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.m] Calculate and interpret an updated probability using Bayes' formula

Answer:

P(outperform)
= 30%

P(stock goes up|outperform)
= 70%

P(outperform and
stock goes up)
= $70\% \times 30\% = 21\%$

P(stock goes down|outperform)
= 30%

P(outperform and
stock goes down)
= $30\% \times 30\% = 9\%$

P(meet
expectation)
= 40%

P(stock goes up|meet expectation)
= 50%

P(meet expectation
and stock goes up)
= $50\% \times 40\% = 20\%$

P(stock goes down|meet expectation)
= 50%

P(meet expectation
and stock goes down)
= $50\% \times 40\% = 20\%$

P(underperform)
= 30%

P(stock goes up|underperform)
= 20%

P(underperform and
stock goes up)
= $20\% \times 30\% = 6\%$

P(stock goes down|underperform)
= 80%

P(underperform and
stock goes down)
= $80\% \times 30\% = 24\%$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.m] Calculate and interpret an updated probability using Bayes' formula

Answer:

We need to calculate the probability that the economy outperformed given that the stock goes up:

$$P(\text{outperform}|\text{stock goes up}) = \frac{P(\text{stock goes up}|\text{outperform})}{P(\text{stock goes up})} \times P(\text{outperform})$$

Step 1: Calculate $P(\text{stock goes up})$

$$P(\text{stock goes up}) = 21\% + 20\% + 6\% = 47\%$$

Step 2: $P(\text{outperform}|\text{stock goes up})$

$$P(\text{outperform}|\text{stock goes up}) = \frac{70\%}{47\%} \times 30\% \approx \mathbf{44.68\%}$$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.n] Identify the most appropriate method to solve a particular counting problem and analyze counting problems using factorial, combination, and permutation concepts

1.

Main principles of counting

Concepts	Definition	Formula
Counting for factorial	The number of ways to assign n different labels to n items	$n! = n(n-1)(n-2)...1$ (n factorial)
Labeling problems (multinomial formula)	The situation where there are n items that can each receive one of k different labels $n_1 + n_2 + \dots + n_k = n$	$\frac{n!}{(n_1!) \times (n_2!) \times \dots \times (n_k!)}$
Combination (binomial formula)	<ul style="list-style-type: none"> General formula for labeling when $k = 2$ Selecting r items from a set of n items when the order of selection is not important 	${}_nC_r = \frac{n!}{(n-r)!r!}$
Permutation	<ul style="list-style-type: none"> A specific ordering of a group of objects The number of different groups of size r in specific order can be chosen from n objects 	${}_nP_r = \frac{n!}{(n-r)!}$

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.n] Identify the most appropriate method to solve a particular counting problem and analyze counting problems using factorial, combination, and permutation concepts

2.

Example

Factorial

Example: Suppose you want to assign three security analysts to cover three different industries. In how many ways can the assignments be made?

Answer:

The number of ways which the assignments can be made are $3! = 6$.

Labelling

Example: A firm is going to create three teams of four from twelve employees. How many ways can the twelve employees be selected for the three teams?

Answer:

This problem is a labeling problem where the 12 employees will be assigned one of three labels. With $n = 12$, $k = 3$, we apply the labeling formula:

$$\frac{12!}{(4!) \times (4!) \times (4!)} = 34,650$$

There are 34,650 ways to group the employees.

MODULE 3: PROBABILITY CONCEPTS

[LOS 3.n] Identify the most appropriate method to solve a particular counting problem and analyze counting problems using factorial, combination, and permutation concepts

2.

Example

Combination

Example: A firm wants to select a team of 5 from a group of 10 employees. How many ways can the firm compose the team of five ?

Answer:

This is a choose a team of 5 from a group of 10 where *order is not a major concern*. With $n = 10$ and $r = 5$, we apply the combination formula:

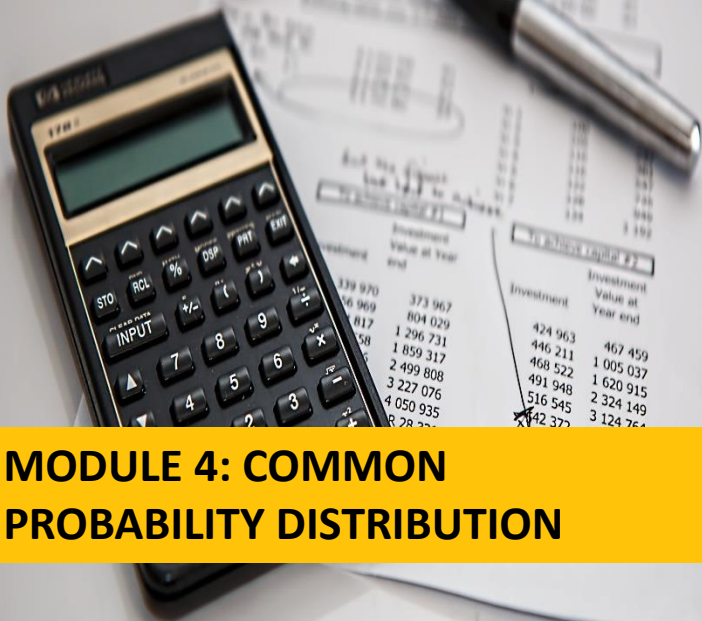
$${}_{10}C_5 = \frac{10!}{(10 - 5)!5!} = \frac{3,628,800}{120 \times 120} = 252.$$

Permutation

Example: A portfolio manager wants to eliminate four stocks from a portfolio that consists of six stocks. How many ways can the four stocks be sold when the order of the sales is important?

Answer:

This is a choose four from six problem where *order is important*. With $n = 6$ and $r = 4$, we apply the permutation formula: ${}_6P_4 = \frac{6!}{(6 - 4)!} = 360.$



MODULE 4: COMMON PROBABILITY DISTRIBUTION

MODULE 4: COMMON PROBABILITY DISTRIBUTION

Learning outcome statements

[LOS 4.a] define a probability distribution and compare and contrast discrete and continuous random variables and their probability functions

[LOS 4.b] calculate and interpret probabilities for a random variable given its cumulative distribution function

[LOS 4.c] describe the properties of a discrete uniform random variable, and calculate and interpret probabilities given the discrete uniform distribution function

[LOS 4.d] describe the properties of the continuous uniform distribution, and calculate and interpret probabilities given a continuous uniform distribution

[LOS 4.e] describe the properties of a Bernoulli random variable and a binomial random variable, and calculate and interpret probabilities given the binomial distribution function

[LOS 4.f] explain the key properties of the normal distribution

[LOS 4.g] contrast a multivariate distribution and a univariate distribution, and explain the role of correlation in the multivariate normal distribution;

[LOS 4.h] calculate the probability that a normally distributed random variable lies inside a given interval;

MODULE 4: COMMON PROBABILITY DISTRIBUTION

Learning outcome statements

[LOS 4.i] explain how to standardize a random variable

[LOS 4.j] calculate and interpret probabilities using the standard normal distribution

[LOS 4.k] define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy's safety-first criterion

[LOS 4.l] explain the relationship between normal and lognormal distributions and why the lognormal distribution is used to model asset prices

[LOS 4.m] calculate and interpret a continuously compounded rate of return, given a specific holding period return

[LOS 4.n] describe the properties of the Student's t-distribution, and calculate and interpret its degrees of freedom

[LOS 4.o] describe the properties of the chi-square distribution and the F-distribution, and calculate and interpret their degrees of freedom

[LOS 4.p] describe Monte Carlo simulation

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.a]: Define a probability distribution and compare and contrast discrete and continuous random variables and their probability functions

1.

Definition of probability function

A probability distribution specifies the probabilities of the possible outcomes of a **random variable**.

2.

Random variable

A random variable is a quantity whose future outcomes are uncertain.

Discrete random variable

A discrete random variable is one for which the number of possible outcomes can be *counted*, and for each possible outcome, there is a measurable and positive probability.

Example:

The number of days it will rain in a given month.

Continuous random variable

A continuous random variable is one for which the number of possible outcomes is *infinite*, even if lower and upper bounds exist.

Example:

The actual amount of daily rainfall between zero and 100 inches.

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.a]: Define a probability distribution and compare and contrast discrete and continuous random variables and their probability functions

3.

Interpretation of probability distributions

A probability distribution is interpreted through either a **probability function** (also known as density function for continuous random variable) or a **cumulative distribution function**.

3.1. Definition of Probability function

A probability function, denoted $p(x)$, is the probability that random variable X takes on the value x , or $p(x) = P(X = x)$.

- $0 \leq p(x) \leq 1$
- $\sum p(x) = 1$

For discrete random variable

$p(x) = 0$ when x cannot occur, or
 $p(x) > 0$ if it can.

For continuous random variable

- $p(x) = 0$ even though x can occur.
- We can only consider $P(x_1 \leq X \leq x_2)$ where x_1 and x_2 are actual numbers.
- $P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2)$ because $P(x_1) = P(x_2) = 0$.

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.b]: Calculate and interpret probabilities for a random variable given its cumulative distribution function

3.2. Definition of Cumulative distribution function

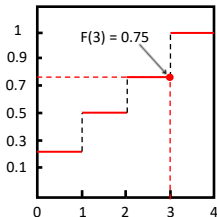
- **Cumulative distribution function (cdf)** defines the probability that a random variable, X , takes on a value *equal to or less than* a specific value, x .
- For both discrete and continuous variable, the notation is $F(x) = P(X \leq x)$.

Example: A random discrete variable X can take one of these values $\{1, 2, 3, 4\}$. $P(1) = P(2) = P(3) = P(4) = 0.25$. Calculate $F(3)$?

Answer:

$$F(3) = P(x \leq 3) = P(1) + P(2) + P(3) = 0.25 + 0.25 + 0.25 = 0.75$$

Illustration of cumulative distribution function (CDF) of a discrete variable



MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.c]: Describe the probabilities of a discrete uniform random variable, and calculate and interpret probabilities given the discrete uniform distribution function

Basic features of discrete uniform distribution

- Uniform distributions are probability distributions with equally likely outcomes.
- In a discrete uniform distribution, outcomes are discrete and have the same probability.

Example:

Given $X = \{1, 2, 3, 4\}$, we have the following probability distributions table:

$X = x$	Probability of x Prob ($X = x$)	Cumulative Distribution Function Prob ($X \leq x$)
1	0.25	0.25
2	0.25	0.50
3	0.25	0.75
4	0.25	1.00

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.d]: Describe the properties of the continuous uniform distribution and calculate and interpret probabilities given a continuous uniform distribution

1.

Basic features of continuous uniform distribution

- Uniform distributions are probability distributions with equally likely outcomes.
- The continuous uniform distribution is defined over a range that spans between some lower limit, a , and some upper limit, b , which serve as the parameters of the distribution.
- Outcomes can only occur between a and b .

2.

Continuous uniform distribution function

Probability density function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Cumulative distribution function

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b \end{cases}$$

Example 1: If variable x has continuous uniform distribution and $0 < x < 10$. Calculate $F(4)$?

Answer:

$$f(x) = \frac{1}{10-0} = \frac{1}{10} \text{ or } 0.1$$

$$F(4) = P(X \leq 4) = p(1) + p(2) + p(3) + p(4) = \frac{4-0}{10-0} = 0.4$$

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.d]: Describe the properties of the continuous uniform distribution and calculate and interpret probabilities given a continuous uniform distribution

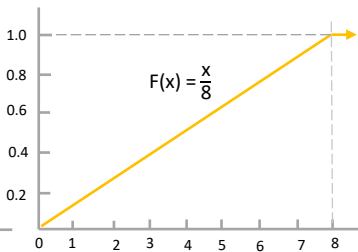
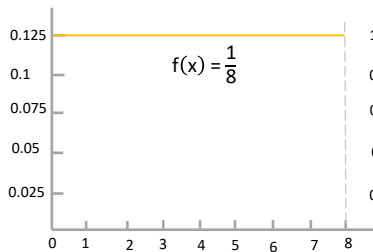
Example 2: Given $x \in (0,8)$, we have the following probability density function and cumulative distribution function

Probability density function

$$f(x) = \begin{cases} \frac{1}{8} & \text{for } 0 \leq x \leq 8 \\ 0 & \text{otherwise} \end{cases}$$

Cumulative distribution function

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{x-0}{8} & \text{for } 0 \leq x \leq 8 \\ 1 & \text{for } x > 8 \end{cases}$$



MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.e]: Describe the properties of a Bernoulli random variable and a binomial random variable, and calculate and interpret probabilities given the binomial distribution function

1.

Basic characteristics of binomial distribution

Definition:

- A Bernoulli trial is an experiment with two outcomes “failure” and “success”.
- The trials are independent, with the probability of success is constantly p .
- **Bernoulli random variable**: variable Y where
 - $Y = 1 \rightarrow$ outcome is success; $p(1) = p$
 - $Y = 0 \rightarrow$ out come is failure; $p(0) = 1 - p$
- **Binomial random variable**: variable X where X is the number of successes in n Bernoulli trials

$$X = Y_1 + Y_2 + \dots + Y_n$$

Formula of binomial probability distribution:

The probability of x successes in n trials:

$$p(x) = p(X = x) = \frac{n!}{(n-x)!x!} \times p^x \times (1-p)^{n-x}$$

(when $p = 0.5$, the binomial distribution is symmetric)

Basic property:

- Expected value of a binomial variable X : $E(X) = n \times p$
- Variance of a binomial variable X : $\sigma_X^2 = n \times p \times (1-p)$

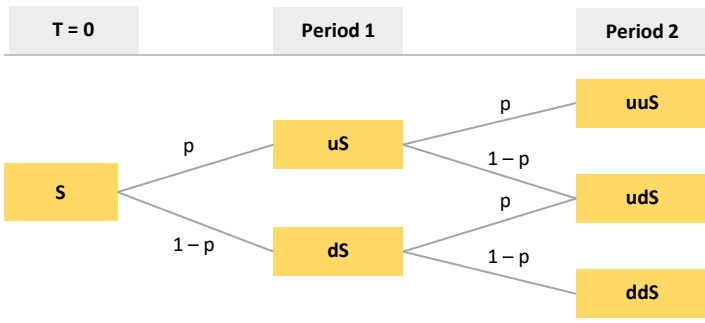
MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.e]: Describe the properties of a Bernoulli random variable and a binomial random variable, and calculate and interpret probabilities given the binomial distribution function

2.

Definition of binomial tree

A binomial tree is the graphical representation of a model of asset price dynamics in which at each period, the asset moves up with probability p or down with probability $(1 - p)$.



MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.e]: Describe the properties of a Bernoulli random variable and a binomial random variable, and calculate and interpret probabilities given the binomial distribution function

2.

Definition of binomial tree

Example:

A stock priced at \$10 has a 60% probability of moving up and a 40% probability of moving down. If it moves up, it increases by a factor of 1.06. If it moves down, it decreases by a factor of $1/1.06$. What is the expected stock price after two successive periods?

MODULE 4: COMMON PROBABILITY DISTRIBUTION

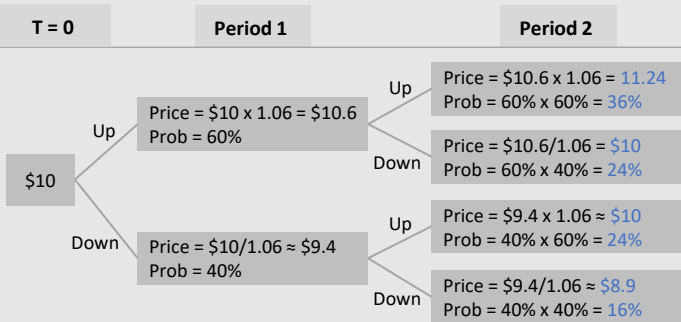
[LOS 4.e]: Describe the properties of a Bernoulli random variable and a binomial random variable, and calculate and interpret probabilities given the binomial distribution function

2.

Definition of binomial tree

Answer:

Step 1: Calculate probability and stock price based on up-down movements of stock price



Step 2: Calculate expected stock price using weighted mean

Expected price = $36\% \times 11.24 + 24\% \times 10 + 24\% \times 10 + 16\% \times 8.9 = \10.27

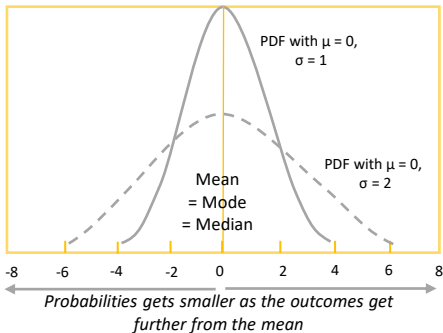
MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.f]: Explain key properties of normal distribution

Key properties of normal distribution

- It is completely described by its mean, μ and variance, σ^2 : $X \sim N(\mu, \sigma^2)$.
- Skewness = 0, meaning that the normal distribution is symmetric about its mean.
- Kurtosis = 3 and excess kurtosis = 0.
- A linear combination of normally distributed random variables is also normally distributed.
- The probabilities of outcomes further above and below the mean get smaller and smaller but do not go to zero (the tails get very thin but extend infinitely).

Normal distribution



MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.g]: Contrast a multivariate distribution and a univariate distribution, and explain the role of correlation in the multivariate normal distribution

1.

Difference between univariate and multivariate distributions

Univariate distribution

A univariate distribution describes a **single random variable**

Multivariate distribution

A multivariate distribution describes the probabilities **for a group of related random variables**.

2.

Key characteristics of multivariate distributions

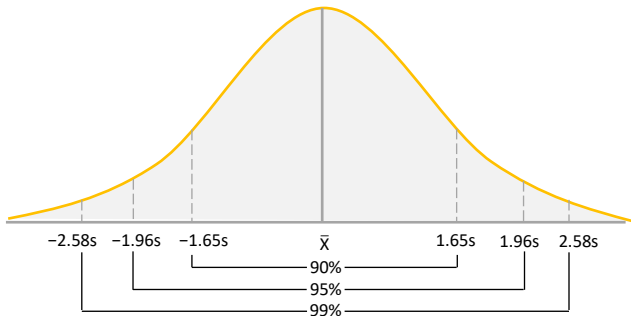
- A multivariate distribution is meaningful when the behavior of each random variable in the group is in some way *dependent* upon the behavior of the others.
- Using asset returns random variables, the multivariate normal distribution for the returns on n assets can be completely defined by the following three sets of parameters:
 - n means of the n series of returns ($\mu_1, \mu_2, \dots, \mu_n$).
 - n variances of the n series of returns ($\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$).
 - $0.5n(n-1)$ pair-wise correlations.

Example: If there are two assets, $n = 2$, then the multivariate returns distribution can be described with two means, two variances, and one correlation [$0.5(2)(2-1) = 1$].

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.h]: Calculate the probability that a normally distributed random variable lies inside a given interval

Confidence intervals



Definition of confidence intervals: A confidence interval is a range of values around the expected outcome within which we expect the actual outcome to be some specified percentage of time.

Common confidence interval for normal distribution:

- The 90% confidence interval for X is $\bar{X} - 1.65s$ to $\bar{X} + 1.65s$
- The 95% confidence interval for X is $\bar{X} - 1.96s$ to $\bar{X} + 1.96s$
- The 99% confidence interval for X is $\bar{X} - 2.58s$ to $\bar{X} + 2.58s$

MODULE 4: COMMON PROBABILITY DISTRIBUTION

**[LOS 4.i]: Explain how to standardize a random variable &
[LOS 4.j]: Calculate and interpret probabilities using the
standard normal distribution**

1.

Definition of standard normal distribution

Through the process of standardizing, probability statements for **any** kind of normally distributed random variable can be made by referring to a **standard normal distribution curve**, which is a normal distribution that has been *standardized* so that it has a mean of 0 and a standard deviation of 1.

2.

Standardizing a random variable (X)

- Standardization is the process of converting an observed value for a random variable to **its z-value**.
- The z-value represents the number of standard deviations away from the population mean, a given observation lies.
- To standardize an observation from a given normal random variable (X), the *z-value* of the observation (Z) must be calculated:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- $X \sim N(\mu, \sigma^2)$
- $Z \sim N(0, 1)$

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.i]: Explain how to standardize a random variable & [LOS 4.j]: Calculate and interpret probabilities using the standard normal distribution

3.**Calculating probabilities using standard normal distribution**

The following procedure describes how to calculate probabilities using z-value:

Have the value
of mean (μ)
and standard
deviation (σ)



Determine
z-value



Use z-table



Determine
probability

Example 1: Considering again EPS distributed with $\mu = \$6$ and $\sigma = \$2$, what is the probability that EPS will be greater than \$9.70?

Step 1: Determine z-value

The z-value must be calculated using the formula:

$$Z = \frac{X - \mu}{\sigma}$$

Answer:

The z-value for EPS = \$9.70 is: $Z = \frac{x - \mu}{\sigma} = \frac{9.7 - 6}{2} = 1.85$

→ *Interpretation:*

$z = +1.85$ indicates that an EPS of \$9.70 is 1.85 standard deviations above the mean EPS value of \$6.

MODULE 4: COMMON PROBABILITY DISTRIBUTION

**[LOS 4.i]: Explain how to standardize a random variable &
[LOS 4.j]: Calculate and interpret probabilities using the
standard normal distribution**

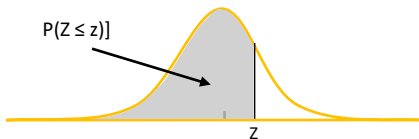
3.

Calculating probabilities using standard normal distribution

Step 2: Determine probability

The values in the z-table are the probabilities of observing a z-value that is equal or less than a given value, z. [$P(Z \leq z)$]

Cdf values for the standard normal distribution: The z-table for $z \geq 0$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Answer (cont.)

From the z-table we have: $F(1.85) = 0.9678$, but this is $P(\text{EPS} \leq 9.70)$.

We want $P(\text{EPS} > 9.70)$, which is $1 - P(\text{EPS} \leq 9.70)$

→ $P(\text{EPS} > 9.70) = 1 - 0.9678 = 0.0322$, or 3.22%.

MODULE 4: COMMON PROBABILITY DISTRIBUTION

**[LOS 4.i]: Explain how to standardize a random variable &
[LOS 4.j]: Calculate and interpret probabilities using the
standard normal distribution**

3.

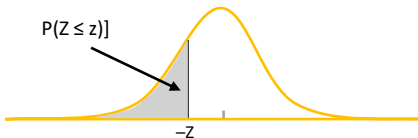
Calculating probabilities using standard normal distribution

Example 2: Continue with previous example, which is $\mu = \$6$ and $\sigma = \$2$, we determine that z-value equals 1.85, but in this case we only have negative z-table. What is the probability that EPS will be greater than \$9.70?

Answer:

In this case we use z-table for $z \leq 0$.

Cdf values for the standard normal distribution: The z-table for $z \leq 0$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

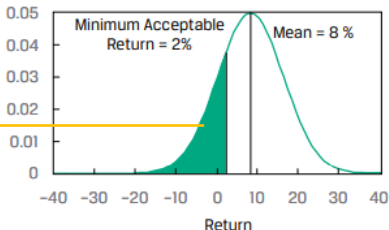
- From the z-table we have $F(-1.85) = 0.0322$.
- $F(-1.85) + F(1.85) = 1 \rightarrow F(1.85) = 1 - F(-1.85) = 0.9678$
 $\rightarrow P(\text{EPS} > 9.70) = 1 - F(1.85) = 1 - 0.9678 = 0.0322$, or 3.22%.

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.k]: Define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy's safety-first criterion

1. Shortfall risk

Shortfall risk is the probability that a portfolio value or return will fall below a particular target (or a minimum acceptance return) over a given period.



2. Roy's safety first criterion

Assumption:

- Returns are normally distributed.
- Investors are risk averse and rational.

Principle: Optimal portfolio is the portfolio that minimizes the shortfall risk or maximizes Roy's safety-first ratio (SFR), where:

$$SFR = \frac{E(R_p) - R_L}{\sigma_p}$$

(R_L : threshold return level, which is the minimum acceptable level)

Interpretation: The higher SFR, the more optimal a portfolio is.

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.k]: Define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy's safety-first criterion

Example: Three portfolios with normally distributed returns are available to an investor who wants to minimize the probability that the portfolio return will be less than 5%. The risk and return characteristics of these portfolios are shown in the following table:

Portfolio	Expected return	Standard deviation
Epps	6%	4%
Flake	7%	9%
Grant	10%	15%

Based on Roy's safety-first criterion, which portfolio should the investor select?

Answer:

Step 1: Calculate each portfolio safety-first ratio

We have $R_L = 5\%$, applying Roy's safety-first ratio for the three portfolios:

- $SFR_{Epps} = \frac{E(R_p) - R_L}{\sigma_p} = \frac{6\% - 5\%}{4\%} = 0.25$
- $SFR_{Flake} = \frac{7\% - 5\%}{9\%} \approx 0.22$
- $SFR_{Grant} = \frac{10\% - 5\%}{15\%} \approx 0.33$

Step 2: Select the portfolio with the highest safety-first ratio

→ The investor should select the Grant portfolio

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.I]: Explain the relationship between normal and lognormal distribution and why the lognormal distribution is used to model asset prices

Lognormal distribution

Definition: The lognormal distribution is generated by the function, e^x where x is normally distributed.

Usage: With a normal distribution of returns to model asset prices over time, we admit the possibility of asset prices less than zero, which is not realistic. So, lognormal distribution is used to solve this problem.

Key characteristics:

- The lognormal distribution is skewed to the right.
- The lognormal distribution is bounded from below by zero so that it is useful for modeling asset prices which never take negative values.

Illustration of Lognormal distribution



MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.I]: Explain the relationship between normal and lognormal distribution and why the lognormal distribution is used to model asset prices

Why the lognormal distribution is used to model asset prices

The stock price at some future time T (S_t) is calculated as:

$$S_t = S_0 \times e^{r(0,T)}$$

where:

- S_0 is the current stock price
- $r(0,T)$ is the continuously compounded return from 0 to T

We can write $r(0,T)$ as the sum of shorter-term continuously compounded returns and that if these shorter-period returns are normally distributed, then $r(0,T)$ is normally distributed (given certain assumptions) or approximately normally distributed (not making those assumptions).

As S_t is proportional to the log of a normal random variable
→ S_t is lognormally distributed.

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.m]: Calculate and interpret a continuously compounded rate of return, given a specific holding period

Distinguish between Continuous and Discrete compounding rate of return

Continuous compounding views time as essentially or unbroken.

Discrete compounding views time as advancing in discrete finite intervals such as monthly or semiannually.

Formula of continuous compounding

- With continuous compounding, effective annual rate = $\text{EAR} = e^{R_{CC}} - 1$
- By reorganizing the EAR equation, we have the continuously compounded states **annual** rate formula for 2 consecutive periods:

$$R_{CC} = \ln\left(\frac{S_1}{S_0}\right) = \ln(1 + \text{EAR})$$

- The relationship between holding period return (HPR) and the continuously compounded rate: $\text{HPR}_t = e^{R_{CC} \times t} - 1 \rightarrow R_{CC} = \frac{\ln(1 + \text{HPR}_t)}{t}$

Example: A stock was purchased for \$100 and sold one year later for \$120. Calculate the investor's annual rate of return on a continuously compounded basis.

Answer:

With $S_0 = \$100$, $S_1 = \$120$, we apply the continuously compounded formula:

$$R_{CC} = \ln\left(\frac{S_1}{S_0}\right) = \ln\left(\frac{120}{100}\right) = 18.23\%$$

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.n]: Describe properties of the Student's t-distribution, and calculate and interpret its degrees of freedom

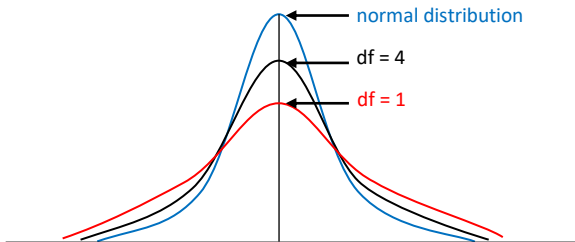
Definition of student's t-distribution

The standard t-distribution is a symmetrical probability distribution defined by a single parameter known as **degrees of freedom (df)**, the number of independent variables used in defining sample statistics, such as variance, and the probability distributions they measure.

T-distribution versus normal distribution

- T-distributions have fatter tail than normal distribution.
- $df = n - 1$ (n is sample size).
- As df increases, the shape of t-distribution approaches normal distribution.

T-distributions for different degrees of freedom (df)



MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.o]: Describe the properties of the chi-square distribution and F-distribution and calculate and interpret their degrees of freedom

Definition of chi-square distribution

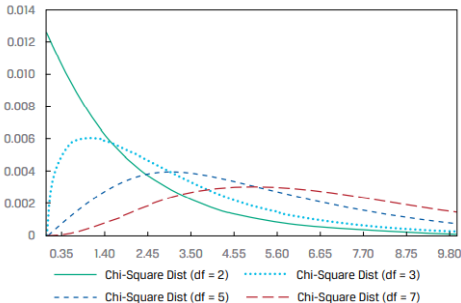
Chi square distribution with k df is the sum of the squares of k independent standard normally distributed random variables.

Illustration

If Z_1, Z_2, \dots, Z_k are independent standard normal random variables, then $(z_1^2 + z_2^2 + \dots + z_k^2)$ has a χ^2 distribution with k degrees of freedom.

Key properties of chi square distribution

- Its shape is asymmetrical (positively skewed).
- It does not take on negative values.



MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.o]: Describe the properties of the chi-square distribution and F-distribution and calculate and interpret their degrees of freedom

Definition of F distribution

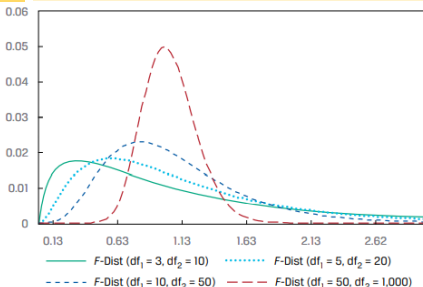
The F-distribution is a family of asymmetrical distributions bounded from below by 0 and directly related to Chi-square distribution.

Illustration

If χ_1^2 is one chi-square random variable with m degrees of freedom and χ_2^2 is another chi-square random variable with n degrees of freedom $\rightarrow F = \frac{\chi_1^2 / m}{\chi_2^2 / n}$ follows an F distribution with m numerator and n denominator degrees of freedom.

Key properties of F distribution

As both the numerator (df_1) and the denominator (df_2) degrees of freedom increase, the density function will also become more bell curve-like.



MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.p]: Describe Monte Carlo simulation

1.

Introduction and general mechanism

- Monte Carlo simulation is a technique based on the repeated generation of one or more risk factors that affect security values, in order to generate a **distribution** of security values.
- A characteristic feature of Monte Carlo simulation is the generation of a large number of random samples from specified probability distributions to represent the operation of risk in the system.

2.

Usage of Monte Carlo Simulation

- Value **complex** securities.
- Simulate the profits/losses from a trading strategy.
- Calculate estimates of **value at risk (VaR)** to determine the riskiness of a portfolio of assets and liabilities.
- Simulate **pension fund assets and liabilities** over time to examine the variability of the difference between the two.
- Value portfolios of assets that have **nonnormal** returns distributions.

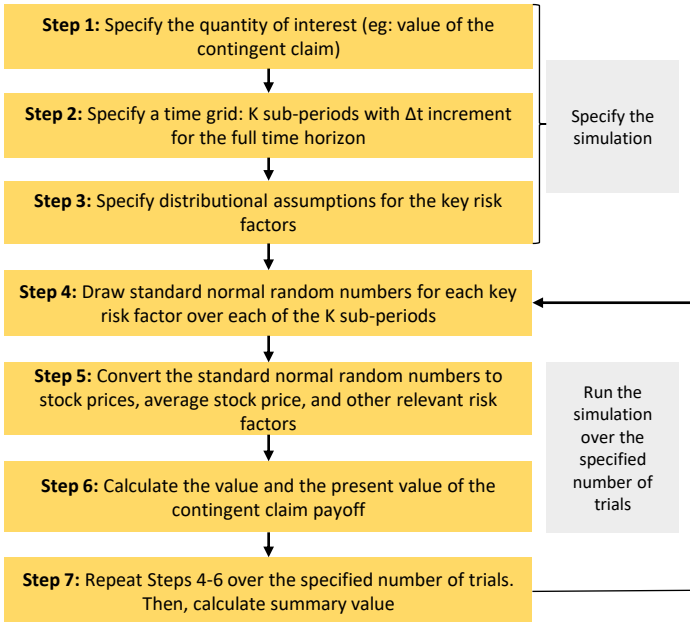
3.

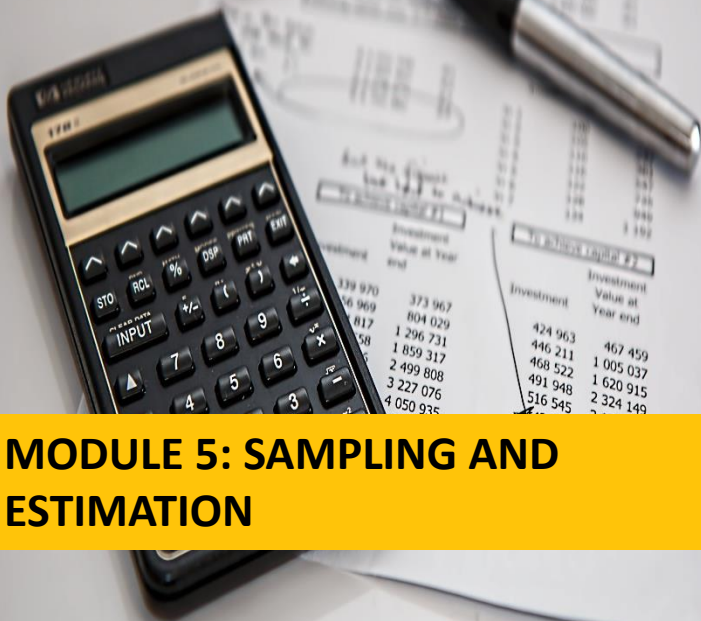
Limitations of Monte Carlo Simulation

- The model is fairly **complex** and will provide answers that are **no better** than the assumptions about the distributions of the risk factors and the pricing/valuation model that is used.
- The model **cannot provide insights** that analytic methods can.

MODULE 4: COMMON PROBABILITY DISTRIBUTION

[LOS 4.p]: Describe Monte Carlo simulation





MODULE 5: SAMPLING AND ESTIMATION

MODULE 5: SAMPLING AND ESTIMATION

Learning outcome statements

[LOS 5.a] compare and contrast probability samples with non-probability samples and discuss applications of each to an investment problem

[LOS 5.b] explain sampling error

[LOS 5.c] compare and contrast simple random, stratified random, cluster, convenience, and judgmental sampling

[LOS 5.d] explain the central limit theorem and its importance

[LOS 5.e] calculate and interpret the standard error of the sample mean

[LOS 5.f] identify and describe desirable properties of an estimator

[LOS 5.g] contrast a point estimate and a confidence interval estimate of a population parameter

[LOS 5.h] calculate and interpret a confidence interval for a population mean, given a normal distribution with 1) a known population variance, 2) an unknown population variance, or 3) an unknown population variance and a large sample size

MODULE 5: SAMPLING AND ESTIMATION

Learning outcome statements

[LOS 5.i] describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

[LOS 5.j] describe the issues regarding selection of the appropriate sample size, data snooping bias, sample selection bias, survivorship bias, look-ahead bias, and time-period bias.

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.a]: Compare and contrast probability samples with non-probability samples and discuss applications of each to an investment problem

There are various methods for obtaining information on a population (all members of a specified group) through samples (part of the population).

Sampling methods

1. Probability sampling

Simple Random Sampling

Systematic Sampling

Stratified Random Sampling

Cluster Sampling

2. Non-probability sampling

Convenience Sampling

Judgment Sampling

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.c]: Compare and contrast simple random, stratified random, cluster, convenience and judgemental sampling

1.

Probability sampling

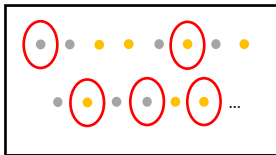
Mechanism: Probability sampling gives every member of the population an **equal chance** of being selected.

Consequence: Samples selected are more representative of population.

Simple random sampling

Mechanism: Each observation is randomly selected without any initial division.

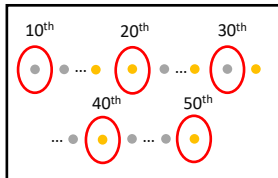
Example: Drawing 5 out of 50 items can be done by numbering each of 50 items and then picking randomly.



Systematic sampling

Mechanism: Every n^{th} item in a population is selected.

Example: Drawing 5 out of 50 items can be done by numbering each of 50 items and then picking the 10th, 20th, 30th, 40th, 50th item.



MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.c]: Compare and contrast simple random, stratified random, cluster, convenience and judgemental sampling

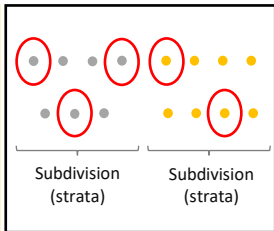
1.

Probability sampling

Stratified random sampling

Mechanism: Population is divided into subgroups, called **strata** or cells, based on certain characteristics, simple random samples are then drawn from each stratum.

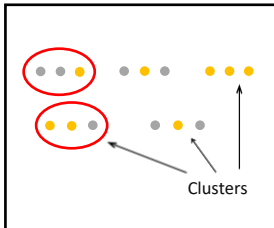
Example: An analyst first divide 10,000 shares into subgroups based on sector then randomly pick shares from each subgroup to form a portfolio.



Cluster sampling

Mechanism: The population is divided into **clusters** (mini representation of the entire population) then certain clusters are chosen as a whole using simple random sampling.

Example: To analyze average income across Vietnam, a research group divided Vietnam into cities and use data from one city or only a few cities.



MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.b]: Explain sampling error [LOS 5.c]: Compare and contrast simple random, stratified random, cluster, convenience and judgemental sampling

2.

Non-probability sampling

Mechanism: Non-probability sampling depends on **factors other than probability** considerations, such as a sampler's judgment or the convenience to access data.

Consequence: Samples selected might not represent population.

Convenience sampling

Mechanism: Selecting sample data process is based on its ease of access, using data that are readily available. Because such a sample is typically not random, sampling error will be greater.

Example: In the preliminary stage of research or in circumstances subject to cost constraints, convenience sampling is often used as a time-efficient and cost-effective sampling plan for a small scale pilot study.

Judgmental sampling

Mechanism: Each observation is selected from a larger data set by the researcher, based on her experience and judgment.

Example: A researcher interested in assessing company compliance with accounting standards may have experience suggesting that evidence of noncompliance is typically found in certain ratios derived from the financial statements. The researcher may select only data on these items.

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.b]: Explain sampling error [LOS 5.c]: Compare and contrast simple random, stratified random, cluster, convenience and judgemental sampling

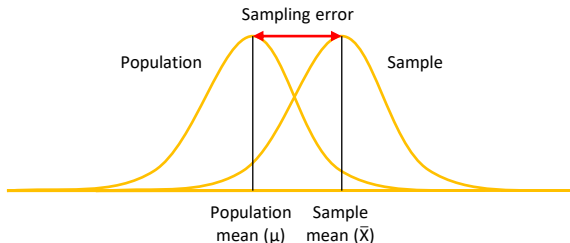
1.

Sampling error

Sampling error is the difference between a sample statistic (*the mean, variance, or standard deviation of the sample*) and its corresponding population parameter (*the true mean, variance, or standard deviation of the population*).

For example, the sampling error for the mean is as follows:

$$\begin{aligned}\text{Sampling error of the mean} &= \text{Sample mean} - \text{Population mean} \\ &= \bar{X} - \mu\end{aligned}$$



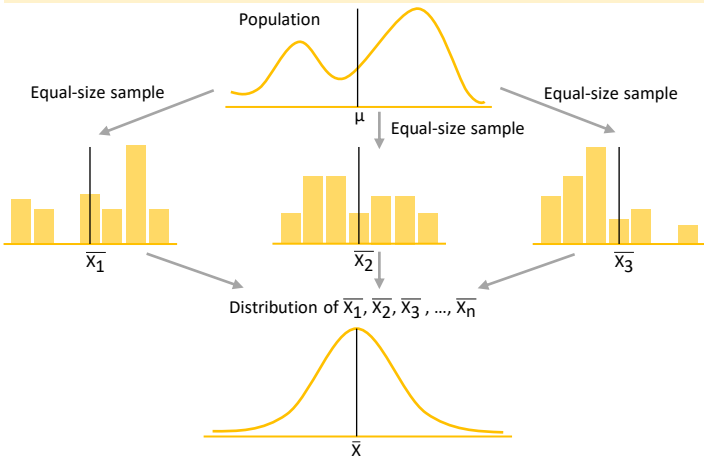
MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.b]: Explain sampling error [LOS 5.c]: Compare and contrast simple random, stratified random, cluster, convenience and judgemental sampling

2.

Sampling distribution

The **sampling distribution** of the sample statistic is a probability distribution of all possible sample statistics computed from a set of *equal-size* samples that were *randomly* drawn from the same population.



MODULE 5: SAMPLING AND ESTIMATION

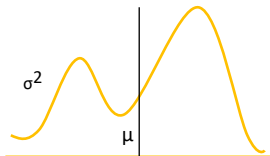
[LOS 5.d]: Explain the central limit theorem and its importance

Central Limit Theorem

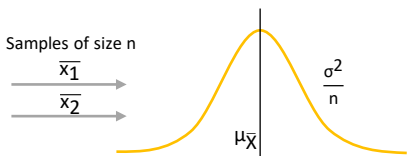
The central limit theorem states that for simple random samples of size n from a *population* with a mean μ and a finite variance σ^2 , the sampling distribution of the sample mean \bar{X} approaches a normal probability distribution with mean μ and a variance equal to $\frac{\sigma^2}{n}$ as the sample size becomes large.

Key properties:

- If the **sample size n is sufficiently large ($n \geq 30$)**, the sampling distribution of the sample means will be **approximately normal**.
- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$



Population Distribution
(does not have to be normal)



Sampling Distribution - Distribution of \bar{X}

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.e]: Calculate and interpret the standard error of the sample mean

Standard error of the sample mean

Definition: The standard error of the sample mean (\bar{X}) is the standard deviation of the distribution of the sample means.

When the population's standard deviation (σ) is **known**

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation

When the population's standard deviation (σ) is **unknown**

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation

Example: Suppose a sample contains the past 30 monthly returns for McCreary, Inc. The sample mean return is 2% and the sample standard deviation is 20%. Calculate and interpret the standard error of the sample mean.

Answer:

Since population's standard deviation (σ) is unknown, the standard error of the sample mean is:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{20\%}{\sqrt{30}} = 3.6\%$$

Interpretation: If we took all possible samples of size 30 from McCreary's monthly returns and prepared a sampling distribution of the sample means, the mean would be 2% with a standard error of 3.6%.

MODULE 5: SAMPLING AND ESTIMATION

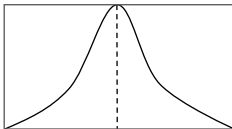
[LOS 5.f]: identify and describe desirable properties of an estimator

Desirable properties of an estimator

Unbiasedness

An unbiased estimator is one whose expected value (the mean of its sampling distribution) equals the parameter it is intended to estimate.

Unbiasedness of an Estimator



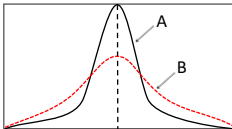
$$E(\bar{X}) = \mu$$

$E(\bar{X}) = \mu \rightarrow$ the sample mean is an unbiased estimator of population mean.

Efficiency

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

Efficiency of an Estimator



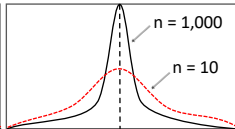
$$E(\bar{X}_A) = E(\bar{X}_B) = \mu$$

- $E(\bar{X}_A) = E(\bar{X}_B) = \mu \rightarrow$ Unbiased estimator
- A shows smaller variance
 \rightarrow A is more efficient

Consistency

A consistent estimator is one for which the probability of estimates close to the value of the population parameter increases as sample size increases.

Consistency of an Estimator



$$E(\bar{X}) = \mu$$

- $E(\bar{X}) = \mu \rightarrow$ Unbiased estimator
- Sample size increases
 \rightarrow Standard error narrows
 \rightarrow More consistent

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.g]: Contrast a point estimate and a confidence interval

1.

Point estimates

Point estimates are single (sample) values used to estimate population parameters. The formula used to compute the point estimate is called the estimator.

Example: The sample mean, \bar{x} , is an estimator of the population mean μ and is computed using the familiar formula:

$$\bar{x} = \frac{\sum x}{n}$$

2.

Confidence intervals

A confidence interval is a range for which one can assert with a given probability $1 - \alpha$ that it will contain the parameter it is intended to estimate.

→ This interval is often referred to as the **100(1 - α)%** confidence interval for the parameter.

- α is called the **level of significance**.
- $1 - \alpha$ is called the **degree of confidence**.

Illustration: A calculated interval range from 15 to 25 at the 5% level of significance implies that we can be 95% confident that the population parameter will lie between 15 and 25.

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.h]: Calculate and interpret a confidence interval for a population mean

Formula to construct confidence intervals

A $100(1 - \alpha)\%$ confidence interval for a parameter has the following structure:

Point estimate \pm (Reliability factor \times Standard error)

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

(Known population variance) (Unknown population variance)

Where:

- Point estimate = a point estimate of the parameter (a value of a sample statistic)
- $z_{\alpha/2}$ or $t_{\alpha/2}$ = Reliability factor: a number based on the assumed distribution of the point estimate and the degree of confidence $(1 - \alpha)$ for the confidence interval.
- α is significance level.
- $\frac{\sigma}{\sqrt{n}}$ = The standard error of the sample mean, where σ is the known standard deviation of the population, and n is the sample size.
- $\frac{s}{\sqrt{n}}$ = The standard error of the sample mean, where s is the standard deviation of the sample when population standard deviation is unknown.

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.h]: Calculate and interpret a confidence interval for a population mean

1.

Confidence intervals for the population mean (normally distributed with **known** population variance)

$$\text{Confidence interval} \in \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The reliability factors for the most frequently used confidence intervals based on the *standard normal distribution* are:

- 90% confidence intervals: Use $z_{0.05} = 1.65$
- 95% confidence intervals: Use $z_{0.025} = 1.96$
- 99% confidence intervals: Use $z_{0.005} = 2.58$

Example: Consider a practice exam that was administered to 36 Level I candidates. The mean score on this practice exam was 80. Assuming a population standard deviation equal to 15, construct and interpret a 99% confidence interval for the mean score on the practice exam for 36 candidates.

Answer:

Note that in this example the population standard deviation is known, so we don't have to estimate it.

At a confidence level of 99%, $\alpha = 1\%$ so $z_{\alpha/2} = z_{0.005} = 2.58$. Thus, the 99% confidence interval is calculated as follows:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 80 \pm 2.58 \frac{15}{\sqrt{36}} = 80 \pm 6.45$$

→ The 99% confidence interval ranges from 73.55 to 86.45.

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.h]: Calculate and interpret a confidence interval for a population mean

2.

Confidence intervals for the population mean (normally distributed with **unknown population variance)**

$$\text{Confidence interval} \in \bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where the number of degrees of freedom is $n - 1$

Unlike the standard normal distribution, the reliability factors for the t -distribution ($t_{\alpha/2}$) depend on the sample size, so we can't rely on a commonly used set of reliability factors.

Example: Suppose a sample contains the past 30 monthly returns for McCreary, Inc. The mean return is 2% and the sample standard deviation is 20%. Find the 95% confidence interval for the mean monthly return.

Answer:

- The T-statistic is used because the population variance is unknown.
- At a confidence level of 95%, $\alpha = 5\%$ so $\alpha/2 = 0.025$ and $df = 30 - 1 = 29$, we use t -table.

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.h]: Calculate and interpret a confidence interval for a population mean

2.

Confidence intervals for the population mean (normally distributed with **unknown population variance)**

The t-table extract from SchweserNotes

Level of significance for one-tailed test						
df	0.100	0.050	0.025	0.01	0.005	0.0005
Level of significance for two-tailed test						
df	0.2	0.1	0.05	0.02	0.01	0.001
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646

The t-table extract from Curriculum

df	0.1	0.05	0.025	0.01	0.005
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750

→ The reliability factor $t_{\alpha/2}$ is: $t_{0.025}^{29} = 2.045$.

Thus, the 95% confidence interval for the population mean is:

$$2\% \pm 2.045 \left(\frac{20\%}{\sqrt{30}} \right) = 2\% \pm 7.4\%$$

→ The 95% confidence interval ranges from -5.4% to 9.4%.

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.h]: Calculate and interpret a confidence interval for a population mean

3.

Criteria for Selecting the appropriate test statistic

When sampling from a:	Test statistic	
	Small sample ($n < 30$)	Large sample ($n \geq 30$)
Normal population with <i>known</i> variance	Z-statistic	Z-statistic
Normal population with <i>unknown</i> variance	T-statistic	T-statistic (*)
Non-normal population with <i>known</i> variance	Not available	Z-statistic
Non-normal population with <i>unknown</i> variance	Not available	T-statistic (*)

* The z-statistic is theoretically acceptable here, but use of t-statistic is more conservative.

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.i]: Describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

Definition of Resampling

Resampling is a computational tool which repeatedly draw samples from the original observe data sample for the statistical inference of population parameters.

Common methods of resampling

Bootstrap

This technique construct the sampling distribution of an estimator by repeatedly drawing samples from the original sample to find standard error and confidence interval.

Jackknife

Unlike bootstrap, which repeatedly draws samples with replacement, jackknife samples are selected by taking the original observed data sample and leaving out one observation at a time from the set (and not replacing it).

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.i]: Describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

Illustration of Bootstrap

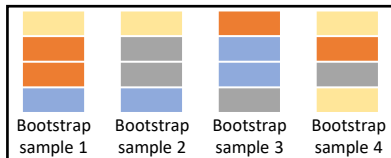
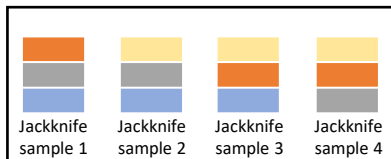


Illustration of Jackknife



Standard deviation of the sample mean in resampling

$$s_{\bar{X}} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b - \bar{\theta})^2}$$

Where

- $s_{\bar{X}}$: the estimate of the standard error of the sample mean
- B : the number of resamples drawn from the original sample
- $\hat{\theta}_b$: the mean of a resample
- $\bar{\theta}$: the mean across all the resample means

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.j]: Describe the issues regarding selection of the appropriate sample size, data snooping bias, sample selection bias, survivorship bias, look-ahead bias, and time period bias

Issues regarding selection of the appropriate sample size

The need for precision

The risk of sampling from a different population

The expenses of different sample sizes

Sampling bias

Data snooping bias

Selection bias

Look-ahead bias

Time period bias

Survivorship bias

Self selection bias

Implicit selection bias

Backfill bias

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.j]: Describe the issues regarding selection of the appropriate sample size, data snooping bias, sample selection bias, survivorship bias, look-ahead bias, and time period bias

Data snooping bias (Data-mining bias)

Data snooping occurs when analysts repeatedly use the same database to search for patterns or trading rules until one that “works” is discovered.

Look ahead bias

Look-ahead bias exists if the model uses data not available to market participants at the time the market participants act in the model.

Time period bias

Time-period bias can result if the time period over which the data is gathered is either too short or too long.

Selection bias

Sample selection bias occurs when data availability leads to certain assets being excluded from the analysis.

- **Survivorship bias** occurs if companies are excluded from the analysis because they have gone out of business or because of reasons related to poor performance.
- **Self-selection bias** reflects the ability of entities to decide whether or not they wish to report their attributes or results and be included in databases or samples.
- **Implicit selection bias** is one type of selection bias introduced through the presence of a threshold that filters out some unqualified members.
- **Backfill bias** occur if past data, not reported or used before, is backfilled into an existing database.

MODULE 5: SAMPLING AND ESTIMATION

[LOS 5.j]: Describe the issues regarding selection of the appropriate sample size, data snooping bias, sample selection bias, survivorship bias, look-ahead bias, and time period bias

Example:

1. A report on long-term stock returns focused exclusively on all currently publicly traded firms in an industry is *most likely* susceptible to:

- A. Look-ahead bias.
- B. Survivorship bias.
- C. Intergenerational data mining.

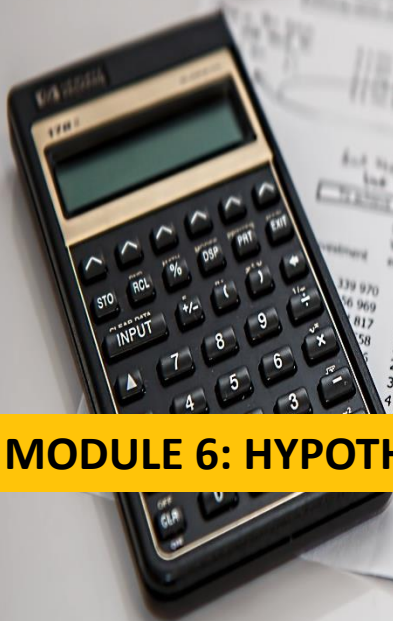
2. Which of the following characteristics of an investment study *most likely* indicates time period bias?

- A. The study is based on a short time-series.
- B. Information not available on the test date is used.
- C. A structural change occurred prior to the start of the study's time series.

Answer:

1. **B** is correct. A report that uses a current list of stocks does not account for firms that failed, merged, or otherwise disappeared from the public equity market in previous years. As a consequence, the report is biased. This type of bias is known as survivorship bias.

2. **A** is correct. A short time series is likely to give period-specific results that may not reflect a longer time period.



MODULE 6: HYPOTHESIS TESTING

MODULE 6: HYPOTHESIS TESTING

Learning outcome statements

LOS 6.a. define a hypothesis, describe the steps of hypothesis testing, and describe and interpret the choice of the null and alternative hypotheses

LOS 6.b. compare and contrast one-tailed and two-tailed tests of hypotheses

LOS 6.c. explain a test statistic, Type I and Type II errors, a significance level, how significance levels are used in hypothesis testing, and the power of a test

LOS 6.d. explain a decision rule and the relation between confidence intervals and hypothesis tests, and determine whether a statistically significant result is also economically meaningful

LOS 6.e. explain and interpret the p-value as it relates to hypothesis testing

LOS 6.f. describe how to interpret the significance of a test in the context of multiple tests;

LOS 6.g. identify the appropriate test statistic and interpret the results for a hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed and the variance is (1) known or (2) unknown

LOS 6.h. identify the appropriate test statistic and interpret the results for a hypothesis test concerning the equality of the population means of two at least approximately normally distributed populations based on independent random samples with equal assumed variances

MODULE 6: HYPOTHESIS TESTING

Learning outcome statements

LOS 6.i. identify the appropriate test statistic and interpret the results for a hypothesis test concerning the mean difference of two normally distributed populations

LOS 6.j. identify the appropriate test statistic and interpret the results for a hypothesis test concerning (1) the variance of a normally distributed population and (2) the equality of the variances of two normally distributed populations based on two independent random samples

LOS 6.k. compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test

LOS 6.l. 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

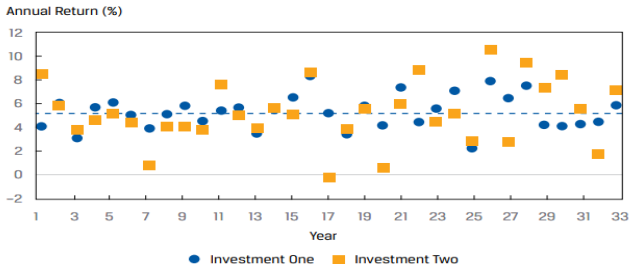
LOS 6.m. explain tests of independence based on contingency table data.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.a]: Define a hypothesis, describe the steps of hypothesis testing, and describe and interpret the choice of the null and alternative analysis

Why we need to use Hypothesis testing?

Returns for investment One and Two over 33 years



The plot above represents the returns on two investments over 33 years, but what can we actually glean from this plot?

- Can we tell if each investment's returns are different from an average of 5%?
- Can we tell whether the returns are different for Investment One and Investment Two?
- Can we tell whether the variability is different for the two investments?

Hypothesis testing is used to address these questions

MODULE 6: HYPOTHESIS TESTING

[LOS 6.a]: Define a hypothesis, describe the steps of hypothesis testing, and describe and interpret the choice of the null and alternative analysis

The process of hypothesis testing

Hypothesis testing is the process of evaluating the accuracy of a statement regarding a population parameter given sample information.

Step 1: State the hypotheses



Step 2: Identify the appropriate test statistic



Step 3: Specify the level of significance



Step 4: State the decision rule



Step 5: Collect data and calculate the test statistic



Step 6: Make a decision

MODULE 6: HYPOTHESIS TESTING

[LOS 6.a]: Define a hypothesis, describe the steps of hypothesis testing, and describe and interpret the choice of the null and alternative analysis

1.

Step 1: State the hypotheses

Null hypothesis (H_0)

Definition: The null hypothesis is the hypothesis that the researcher wants to reject.

H_0 can be stated as (considering population mean):

$$\mu = \mu_0$$

$$\mu \leq \mu_0$$

$$\mu \geq \mu_0$$

Alternative hypothesis (H_a)

Definition: The alternative hypothesis is the hypothesis concluded if H_0 is rejected, also called the hoped-for hypothesis.

H_a can be stated as (considering population mean):

$$\mu \neq \mu_0$$

$$\mu > \mu_0$$

$$\mu < \mu_0$$

Note: The **most common** null hypothesis will include the “equal to” sign and the alternative will include the “not equal to” sign. However, the null hypothesis sometimes can be a “not equal to” hypothesis, combined with an “equal to” alternative.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.a]: Define a hypothesis, describe the steps of hypothesis testing, and describe and interpret the choice of the null and alternative analysis

1.

Step 1: State the hypotheses

Example 1:

State the hypotheses of the test if we want to test whether:

- (i) the population mean return is equal to 6%.
- (ii) the population mean return is greater than 6%.
- (iii) the population mean return is less than 6%.

Answer:

Recall the basis statement:

- The null hypothesis (H_0) is the hypothesis that we are *interested in rejecting*.
- The alternate hypothesis (H_a) is essentially the statement whose *validity we are trying to evaluate*.

(i) We would state the hypotheses as:

$$H_0: \mu = 6$$

$$H_a: \mu \neq 6.$$

(ii) We would state the hypotheses as:

$H_0: \mu \leq 6 \rightarrow$ this is the hypothesis we want to reject.

$H_a: \mu > 6 \rightarrow$ this is the “hope-for” hypothesis, which is also the hypothesis we want to test.

(ii) We would state the hypotheses as:

$$H_0: \mu \geq 6$$

$$H_a: \mu < 6.$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.b]: Compare and contrast one tailed and two tailed test

1.

Step 1: State the hypotheses

One-tailed test

Test structure is either:

- **Right tail:**

$$H_0: \mu \leq \mu_0 \text{ versus } H_a: \mu > \mu_0$$

- **Left tail:**

$$H_0: \mu \geq \mu_0 \text{ versus } H_a: \mu < \mu_0$$

Two-tailed test

Test structure is:

$$H_0: \mu = \mu_0 \text{ versus } H_a: \mu \neq \mu_0$$

Note: The null and alternative hypotheses must be *mutually exclusive* and *collectively exhaustive*; in other words, all possible values are contained in either the null or the alternative hypothesis.

Illustration 2:

Continue with the previous example 1, we classify test structure as follow:

One-tailed test

- Right tail:

$$H_0: \mu \leq 6 \text{ versus } H_a: \mu > 6$$

- Left tail:

$$H_0: \mu \geq 6 \text{ versus } H_a: \mu < 6$$

Two-tailed test

$$H_0: \mu = 6 \text{ versus } H_a: \mu \neq 6$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.c]: Explain a test statistic, type I error and type II errors, how significance levels are used in hypothesis testing, and the power of a test

2.

Step 2: Identify the appropriate test statistic

Definition: A test statistic is a value calculated using a sample, when used in conjunction with a decision rule, is the basis for deciding whether to reject the null hypothesis.

Formula:

$$\text{Test statistic} = \frac{\text{Sample statistic} - \text{Hypothesized value}}{\text{Standard error of the sample statistics}}$$

Illustration 3:

Consider the sample mean (\bar{X}) calculated from a sample of returns drawn from the population. We have test statistic of a population mean in two cases:

When the population's standard deviation (σ) is **known**

$$\text{Test statistic} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

where σ is the population standard deviation

When the population's standard deviation (σ) is **unknown**

$$\text{Test statistic} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

where s is the sample standard deviation

MODULE 6: HYPOTHESIS TESTING

[LOS 6.c]: Explain a test statistic, type I error and type II errors, how significance levels are used in hypothesis testing, and the power of a test

2.

Step 2: Identify the appropriate test statistic

The key to hypothesis testing is identifying the appropriate test statistic for the hypotheses and the underlying distribution of the population.

Test statistic

Distribution of test statistic

t-distribution

z-distribution

F-distribution

Chi-square
distribution

All of these distributions will be further discussed in this MODULE

MODULE 6: HYPOTHESIS TESTING

[LOS 6.c]: Explain a test statistic, type I error and type II errors, how significance levels are used in hypothesis testing, and the power of a test

3.

Step 3: Specify the level of significance

Since there is always a possibility that the sample may not be perfectly representative of the population, resulting in the conclusions drawn from the test may be wrong, there are two types of errors that can be made when conducting a hypothesis test:

- **Type I error:** rejection of the null hypothesis when it is actually true (false positive).
- **Type II error:** failure to reject the null hypothesis when it is actually false (false negative).

→ These errors are mutually exclusive.

Decision	True situation	
	H_0 true	H_0 false
Fail to reject H_0	Correct decision $P = \text{confidence level}$ $= 1 - \alpha$	Incorrect: Type II error $P(\text{type II error}) = \beta$
Reject H_0	Incorrect: Type I error $P(\text{type I error}) = \alpha$ (level of significance)	Correct decision $P = \text{power of the test}$ $= 1 - \beta$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

4.

Step 4: State the decision rule

4.1. Determining the critical value

Concept of decision rule:

- The decision we take is based on comparing the calculated test statistic with a specified value or values, which we refer to as **critical values**.
- The critical value or values we choose are based on the level of significance and the probability distribution associated with the test statistic.

Note:

- If we find that the calculated value of the test statistic is more extreme than the critical value or values, then we reject the null hypothesis; we say the result is **statistically significant**. Otherwise, we fail to reject the null hypothesis; there is not sufficient evidence to reject the null hypothesis.
- It is statistically incorrect to say “accept” the null hypothesis.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

4.

Step 4: State the decision rule

4.1. Determining the critical value

For one-tailed test

• Right tail:

$H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$

Reject H_0 if: test statistic > **upper critical value**

Do not reject H_0 if: test statistic \leq **upper critical value**

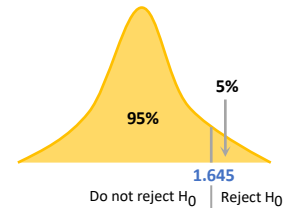
• Left tail:

$H_0: \mu \geq \mu_0$ versus $H_a: \mu < \mu_0$

Reject H_0 if test statistic < **lower critical value**

Do not reject H_0 if: test statistic \geq **lower critical value**

Right one-tailed test using standard normal distribution with $\alpha = 5\%$



For two-tailed test

$H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$

• Reject H_0 if:

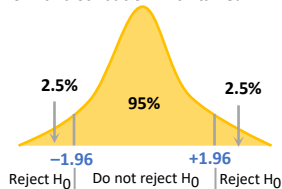
test statistic > **upper critical value** or

test statistic < **lower critical value**

• Do not reject H_0 if:

lower critical value \leq test statistic \leq **upper critical value**

Two-tailed test using standard normal distribution with $\alpha = 5\%$



MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

4.

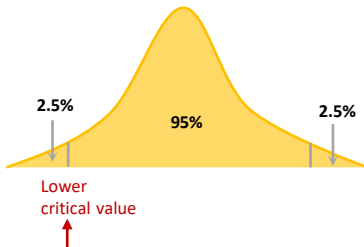
Step 4: State the decision rule

4.1. Determining the critical value

Illustration 4:

Suppose we conduct a two-tailed test at 5% significance level, and the test statistic follows the z-value distribution.

The critical value is therefore determined on the z-distribution at $\alpha/2 = 5\%/2 = 2.5\%$ for both left and right tail.



We look this up in a cumulative z-table ($P(Z) \leq z$) for $z \leq 0$ with a $P(Z) = 2.5\%$

MODULE 6: HYPOTHESIS TESTING

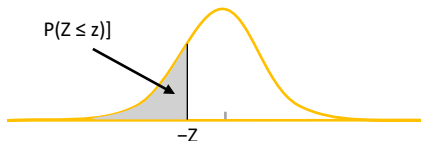
[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

4.

Step 4: State the decision rule

4.1. Determining the critical value

Cumulative z-table with $P(Z \leq z) = N(z)$ for $z \leq 0$



Cdf values for the standard normal distribution: The z-table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

According to the z-table, from $P(Z) = 2.5\%$ we have z-critical = -1.96 , because it is standard normal distribution, so we have z-critical for each tail is:

- Upper z-value = $+1.96$
- Lower z-value = -1.96

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

4.

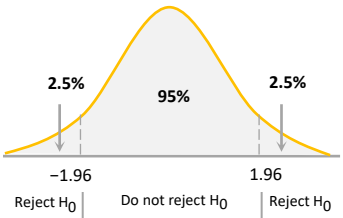
Step 4: State the decision rule

4.2. The relation between confidence intervals and hypothesis tests

Illustration 5:

Suppose we conduct a two-tailed test of a population mean at 5% significance level, and the test statistic follows the z-value distribution.

Hypothesis tests



The conditions for rejecting the null hypothesis is either:

- test statistic > upper critical value

$$\rightarrow \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > 1.96$$

$$\rightarrow \mu_0 < \bar{X} - 1.96 \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

- test statistic < lower critical value

$$\rightarrow \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -1.96$$

$$\rightarrow \mu_0 > \bar{X} + 1.96 \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

MODULE 6: HYPOTHESIS TESTING

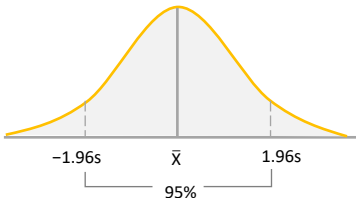
[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

4.

Step 4: State the decision rule

4.2. The relation between confidence intervals and hypothesis tests

Confidence intervals



The 95% confidence interval for the population mean (μ) based on sample mean (\bar{X}) is:

$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

Conclusion:

We reach a similar result between hypothesis test and confidence intervals: We can make a decision by either comparing the calculated test statistic with the critical values or comparing the hypothesized population parameter ($\mu = \mu_0$) with the bounds of the confidence interval.

A significance level in a two-sided hypothesis test can be interpreted in the same way as a $(1 - \alpha)$ confidence interval.

MODULE 6: HYPOTHESIS TESTING

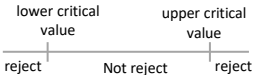
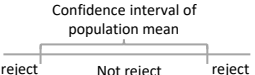
[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

4.

Step 4: State the decision rule

4.3. Decision rules

Making a decision based on Critical values and Confidence intervals for a two-tailed hypothesis:

	Method 1: Based on critical values	Method 2: Based on confidence intervals
Procedure	Compare the calculated test statistic with the critical values .	Compare the calculated test statistic with the bounds of the confidence interval .
Decision	If the calculated test statistic is less than the lower critical value or greater than the upper critical value, reject the null hypothesis.	If the hypothesized value of the population parameter under the null is outside the corresponding confidence interval, the null hypothesis is rejected.
Illustration		

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

5.

Step 5: Collect data and calculate the test statistic

Since we state the appropriate test statistics and their distributions in step 2, in this step, we calculate the formula of test statistics.

Example 6:

Suppose that the basketball player's average score in a sample of 49 games is 36 points with a standard deviation of 9 points. Determine the test statistic to test whether his career scoring average is greater than 30 points.

Answer:

- Recalling the formula of test statistic which is stated in Step 2:

$$\text{Test statistic} = \frac{\text{Sample statistic} - \text{Hypothesized value}}{\text{Standard error of the sample statistics}}$$

- Sample statistic = Sample average score = 36
- Hypothesized value = 30
- Standard error = s/\sqrt{n} with s (sample standard deviation) = 9 and $n = 49$

$$\rightarrow \text{Test statistic} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{36 - 30}{9/\sqrt{49}} = 4.67.$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

6.

Step 6: Make a decision

Example of confidence intervals and two-tailed hypothesis tests

Example 7: A researcher has gathered data on the daily returns on a portfolio of call options over a recent 250-day period. The mean daily return has been 0.1%, and the sample standard deviation of daily portfolio returns is 0.25%. The researcher believes that the mean daily portfolio return is not equal to zero.

1. Construct a hypothesis test of the researcher's belief at 5% significance level.
2. Construct a 95% confidence interval for the population mean daily return over the 250-day sample period.

Answer:

1. Approach 1: Using test statistic

Step 1: State the hypotheses

$$H_0: \mu = 0$$

$$H_a: \mu \neq 0$$

Step 2: Identify the appropriate statistic

Here, we use z-distributed test statistic to test whether the population mean is different than 0.

Step 3: Specify the level of significance

$$\alpha = 5\% \text{ (two tail)}$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

6.

Step 6: Make a decision

Example of confidence intervals and two-tailed hypothesis tests

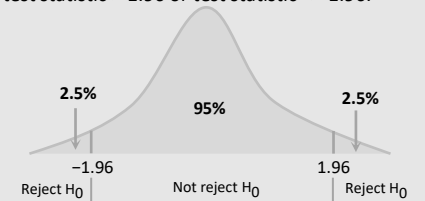
Answer:

1. Approach 1: Using test statistic

Step 4: State the decision rule

At the 5% level of significance of a two-tailed test, the critical z-values for the confidence intervals are $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$.

→ Reject H_0 if test statistic > 1.96 or test statistic < -1.96 .



Step 5: Calculate the test statistic

$$\text{Test statistic} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{0.001 - 0}{0.0025/\sqrt{250}} = 6.325$$

Step 6: Make a decision

Test statistic = 6.325 $>$ critical z-values = 1.96

→ Reject H_0 → The mean daily portfolio return is not equal to zero.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

6.

Step 6: Make a decision

Example of confidence intervals and two-tailed hypothesis tests

Answer:

2. Approach 2: Using confidence intervals

- As the population standard deviation is *unknown* and the sample size is *large* ($n \geq 30$), we can use z-statistic to construct the 95% confidence interval for the population mean daily return over the 250-day sample period is calculated as: $\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

Or we can say that: $\bar{X} - z_{\alpha/2} \times \left(\frac{s}{\sqrt{n}}\right) \leq \mu_0 \leq \bar{X} + z_{\alpha/2} \times \left(\frac{s}{\sqrt{n}}\right)$

- Given a sample size $n = 250$ with a standard deviation $s = 0.25\%$, the standard error can be computed as: $s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.25}{\sqrt{250}} = 0.0158\%$.
 - At the 5% level of significance of a two-tailed test, the critical z-values for the confidence intervals are $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$.
- Given a sample mean $\bar{X} = 0.1\%$, The 95% confidence interval for the population mean is:

$$0.1 - 1.96 \times 0.0158 \leq \mu_0 \leq 0.1 + 1.96 \times 0.0158$$

→ We have $0.069\% \leq \text{population mean daily return} \leq 0.131\%$.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

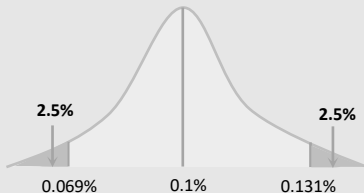
6.

Step 6: Make a decision

Example of confidence intervals and two-tailed hypothesis tests

Answer:

2. Approach 2: Using confidence intervals



In conclusion, there is 95% probability that population mean daily return is ranged from 0.069% to 0.131%, which is different from 0.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

6.

Step 6: Make a decision

Statistical decision

If the calculated test statistic is more extreme than the critical value, statistical decision is that there is sufficient evidence to reject the null hypothesis.

Economic decision

The economic or investment decision takes into consideration not only the statistical decision but also all pertinent economic issues.

Illustration 8: An analyst is testing whether there are positive risk-adjusted returns to a trading strategy. He collects a sample and tests the hypotheses of $H_0: \mu \leq 0\%$ versus $H_a: \mu > 0\%$, where μ is the population mean risk-adjusted return. The mean risk-adjusted return for the sample is 0.7%. The calculated t-statistic is 2.428, and the critical t-value is 2.345. He estimates that the transaction costs are 0.3%.

t-statistic = 2.428 > critical t-values = 2.345.

→ We rejected a null hypothesis and conclude that the mean risk-adjusted return is greater than 0%.

Risk-adjusted return exceeds the transaction cost associated with this strategy by 0.4% (= 0.7 – 0.3).

This result is statistically significant

This result is economically significant

MODULE 6: HYPOTHESIS TESTING

[LOS 6.d]: Explain a decision rule and the relation between confidence intervals, and hypothesis test

6.

Step 6: Make a decision

Statistically significant but not Economically significant

Although a strategy provides a statistically significant, the results may not be economically significant when we account for transaction costs, taxes, and risk.

Illustration 9:

Continue with the illustration 6, but in case that the analyst estimates that the transaction costs are 0.7%.

t-statistic = 2.428 > critical t-values = 2.345.

→ We rejected a null hypothesis and conclude that the mean risk-adjusted return is greater than 0%.

Risk-adjusted return does **not** exceed the transaction cost associated with this strategy as $0.7 - 0.7 = 0\%$.

This result is statistically significant

This result is not economically significant

MODULE 6: HYPOTHESIS TESTING

[LOS 6.e]: Explain and interpret the p-value as it relates to hypothesis testing

Note: Beside hypothesis test, in this LOS we approach another way to test whether H_0 is rejected.

Basic characteristics of p-value

Definition

The p-value is the **smallest level of significance** for which the null hypothesis can be rejected, or we can say that it is the probability of obtaining a test statistic that would lead to a rejection of the null hypothesis.

Position

The p-value is the area in the probability distribution **outside** the calculated test statistic.

Interpretation

The smaller the p-value, the stronger the evidence against the null hypothesis and in favor of the alternative hypothesis.
→ The smaller the chance of making a Type I error.

P-value approach

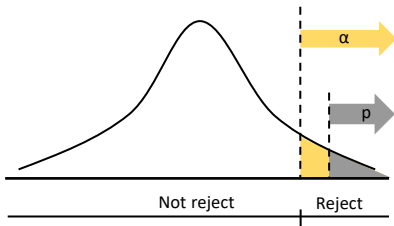
- Reject H_0 when p-value < significance level.
- Do not reject H_0 when p-value > significance level.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.e]: Explain and interpret the p-value as it relates to hypothesis testing

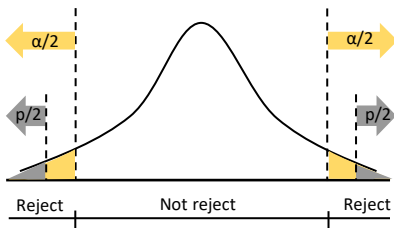
The position of p-value in each type of test

One-tailed test (right tail)



P-value represents probability that lies **above** the computed test statistic.

Two-tailed test



P-value represents probability that lies **above** the positive value of the computed test statistic **plus** the probability that lies **below** the negative value of the computed test statistic.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.e]: Explain and interpret the p-value as it relates to hypothesis testing

Example 10:

An analyst is testing the hypotheses:

$$H_0: \mu = 6$$

$$H_a: \mu \neq 6.$$

Using software, she determines that the p-value for the test statistic is 0.03, or 3%. Make a decision based on p-value in case of:

(i) Level of significance $\alpha = 5\%$.

(ii) Level of significance $\alpha = 1\%$.

Answer:

We have decision based on p-value is: Reject H_0 when p-value < significance level and Do not reject H_0 when p-value > significance level.

Case 1 : Level of significance $\alpha = 5\%$.

p-value = 3% < significant level = 5% \rightarrow Reject H_0 .

Case 2: Level of significance $\alpha = 1\%$.

p-value = 3% > significant level = 1% \rightarrow Do not reject H_0 .

\rightarrow The null is rejected at the 5% level of significance but not at the 1% level of significance.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.f]: Describe how to interpret the significance of a test in the context of multiple tests

Illustration 11:

Suppose we test the hypothesis that the population mean is equal to 6% at 5% significance level and repeat the sampling process by drawing 20 samples and calculating 20 test statistics. We get 6 test statistics with lowest p-values, which are shown in table below.

\bar{X}	0.0664	0.0645	0.0642	0.0642	0.0641	0.0637
Calculated z-statistic	3.1966	2.2463	2.0993	2.0756	2.0723	1.8627
P-value	0.0014	0.0247	0.0358	0.0379	0.0382	0.0625
Reject H_0 ?	Yes	Yes	Yes	Yes	Yes	No

- Using 5% level of significance, if we simply relied on each test and its p-value, then there are 5 tests in which we would reject the null.
- Of the 20 samples tested, we should expect: $20 \times 5\% = 1$ **false positive** (this is known as **type I error** (level of significance), which is the probability that reject the null hypothesis when it is actually true).
- Now we will use False discovery approach (using BH criteria) to adjust p-value to decide the tests in which H_0 is true rejected.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.f]: Describe how to interpret the significance of a test in the context of multiple tests

The process of False discovery approach is presented in the following steps:

Step 1: Rank the p-values from various test from lowest to highest

Step 2: Calculate p-value according to BH criteria:

$$\text{adjusted p-value} = \alpha \times \frac{\text{Rank of p-value}}{\text{Number of tests}}$$

Step 3: Compare original p-value with p-value according to BH criteria.

P-value	Rank	$\alpha \times \frac{\text{Rank of p-value}}{\text{Number of tests}}$	Is p-value \leq adjusted p-value?
0.0014	1	$0.05 \times \frac{1}{20} = 0.0025$	Yes
0.0247	2	$0.05 \times \frac{2}{20} = 0.005$	No
0.0358	3	$0.05 \times \frac{3}{20} = 0.0075$	No
0.0379	4	$0.05 \times \frac{4}{20} = 0.01$	No
0.0382	5	$0.05 \times \frac{5}{20} = 0.0125$	No
0.0625	6	$0.05 \times \frac{6}{20} = 0.015$	No

MODULE 6: HYPOTHESIS TESTING

[LOS 6.f]: Describe how to interpret the significance of a test in the context of multiple tests

Illustration 11 (cont):

According to the table, there is only 1 actual rejection based on comparison of their p-values with their adjusted p-values.

→ The number of significant sample results is the same as would be expected, given the 5% level of significance.

Conclusion:

The results for the samples with Ranks 2 to 5 are false discoveries, and we have not uncovered any evidence from our testing that supports rejecting the null hypothesis, or we can say that, we will not reject the null hypothesis.

In multiple testing (testing more than one time), we can see that if we simply relied on each test and its p-value, then there is a risk of a false discovery.

→ This is referred to as the **multiple testing problem**.

The **false discovery approach** to testing requires using BH criteria in order to avoid misleading (data snooping bias) from multiple testing problems.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.g]: Hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed

Hypothesis test concerning the population mean

Purpose: The purpose of this test is to check whether there is a difference between the population mean and hypothesized value (μ_0).

Step 1: State the hypotheses

- $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$
- $H_0: \mu \geq \mu_0$ versus $H_a: \mu < \mu_0$
- $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$



Step 2: Identify the appropriate test statistic

- This test statistic follows either t-distribution or z-distribution.
- With t-distribution, we use t-statistic.
- With z-distribution, we use z-statistic.

(Condition to use each test is presented in next slide)

T-test

$$T\text{-test} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

with $n - 1$ degrees of freedom

Z-test

$$(1) z\text{-test} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \text{ (large sample)}$$

$$(2) z\text{-test} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \text{ (small sample)}$$

\bar{X} is sample mean

μ_0 is hypothesized population mean

σ is population standard deviation

s is sample standard deviation

MODULE 6: HYPOTHESIS TESTING

[LOS 6.g]: Hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed

Hypothesis test concerning the population mean

Step 2: Identify the appropriate test statistic (cont)

When sampling from a:	Test statistic	
	Small sample ($n < 30$)	Large sample ($n \geq 30$)
Normal population with <i>known</i> variance	Z-statistic	Z-statistic
Normal population with <i>unknown</i> variance	T-statistic	T-statistic (*)
Non-normal population with <i>known</i> variance	Not available	Z-statistic
Non-normal population with <i>unknown</i> variance	Not available	T-statistic (*)

* The z-statistic is theoretically acceptable here, but use of t-statistic is more conservative.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.g]: Hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed

Hypothesis test concerning the population mean

Step 3: Specify the level of significance

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.



Step 4: State the decision rule

- We use appropriate table (z-table or t-table) with consistent significance level to determine critical value.
- Reject H_0 if: test statistic > **upper critical value** or test statistic < **lower critical value**



Step 5: Test statistic

As we state in Step 2, the test statistic can be calculated using t-statistic or z-statistic.



Step 6: Make a decision

We compare calculated Test statistic with critical value and draw appropriate conclusions.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.g]: Hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed

Example 12:

Suppose we want to test whether the daily return in the ACE High Yield Total Return Index is different from zero. Collecting a sample of 1,304 daily returns, we find a mean daily return of 0.0157%, with a sample standard deviation of 0.3157%.

1. Using the t-distributed test statistic to test whether the mean daily return is different from zero at the 5% level of significance.
2. Using the z-distributed test statistic as an approximation, test whether the mean daily return is different from zero at the 5% level of significance.

Answer:

According to the question, as the population standard deviation is *unknown* and sample size is *large* ($n \geq 30$), we can use either z-statistic or t-statistic.

Solution 1: Using t-statistic

Step 1: State the hypotheses

$$H_0: \mu = 0\%$$

$$H_A: \mu \neq 0\%$$

Step 2: Identify the appropriate statistic

Here, we use t-distributed test statistic to test whether the population mean is different than 0% with degrees of freedom = $n - 1 = 1,304 - 1 = 1,303$.

Step 3: Specify the level of significance

$$\alpha = 5\% \text{ (two tail)}$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.g]: Hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed

Answer:

Step 4: State the decision rule

- With $df = 1,303$ we use t-table for two-tailed test with significance level of 5%

The t-table

df	0.1	0.05	0.025	0.01	0.005
120	1.289	1.658	1.980	2.358	2.617
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.960	2.326	2.576

- With $\alpha = 0.05$ and a two-tailed test, probability in each tail would be $\alpha/2 = 0.025$, giving $t_{0.025}^{1,303} \approx 1.960 \rightarrow$ We have t-critical = ± 1.960
 \rightarrow We reject H_0 when Test statistic $> +1.960$ or Test statistic < -1.960 .

Step 5: Calculate the test statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{0.0157 - 0}{0.3157/\sqrt{1,304}} = 1.796$$

Step 6: Make a decision

$$t\text{-statistic} = 1.796 < t\text{-critical} = +1.960$$

\rightarrow Do not reject H_0

\rightarrow The population mean daily return can be equal 0% at 5% level of significance.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.g]: Hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed

Solution 2: Using z-statistic

Step 1: State the hypotheses

$$H_0: \mu = 0\%$$

$$H_a: \mu \neq 0\%$$

Step 2: Identify the appropriate statistic

Here, we use z-distributed test statistic to test whether the population mean is different than 0%.

Step 3: Specify the level of significance

$$\alpha = 5\% \text{ (two tail)}$$

Step 4: State the decision rule

At the 5% level of significance of a two-tailed test, the critical z-values for the confidence intervals are $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$.

→ Reject H_0 if test statistic > 1.96 or test statistic < -1.96 .

Step 5: Calculate the test statistic

$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{0.0157 - 0}{0.3157/\sqrt{1,304}} = 1.796$$

Step 6: Make a decision

$$z\text{-statistic} = 1.796 < z\text{-critical} = +1.960$$

→ Do not reject H_0

→ The population mean daily return can be equal 0% at 5% level of significance.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.h]: Hypothesis test concerning the equality of the population means of two at least approximately normally distributed populations based on independent random samples with equal assumed variances.

Hypothesis test concerning the equality of two population means

Purpose: The purpose of this test is to check if there is a difference between the means of two populations.

Requirement: The samples are required to be *independent* and be taken from two populations that are normally distributed.

Assumption: Our focus in discussing the test of the differences of means is using the assumption that the population variances are *equal*.

Step 1: State the hypotheses



Step 2: Identify the appropriate test statistic

- $H_0: \mu_1 - \mu_2 = 0$ versus $H_a: \mu_1 - \mu_2 \neq 0$
- $H_0: \mu_1 - \mu_2 \geq 0$ versus $H_a: \mu_1 - \mu_2 < 0$
- $H_0: \mu_1 - \mu_2 \leq 0$ versus $H_a: \mu_1 - \mu_2 > 0$

This test statistic follows t-distribution, which is referred to as t-test.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}}$$

with:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$df = n_1 + n_2 - 2$$

where:

s_1^2 : variance of the 1st sample

s_2^2 : variance of the 2nd sample

n_1 : number of observations in the 1st sample

n_2 : number of observations in the 2nd sample

s_p^2 : **pooled estimator** of the common variance

MODULE 6: HYPOTHESIS TESTING

[LOS 6.h]: Hypothesis test concerning the equality of the population means of two at least approximately normally distributed populations based on independent random samples with equal assumed variances.

Hypothesis test concerning the equality of two population means

Step 3: Specify the level of significance



Step 4: State the decision rule



Step 5: Test statistic



Step 6: Make a decision

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.

- We use t-table with consistent significance level to determine t-critical value.
- Reject H_0 if: t-statistic > **upper t-critical** or t-statistic < **lower t-critical**

As we state in Step 2, t-statistic is calculated as:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}}$$

We compare calculated Test statistic with critical value and draw appropriate conclusions.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.h]: Hypothesis test concerning the equality of the population means of two at least approximately normally distributed populations based on independent random samples with equal assumed variances.

Example 13:

Continue with the example 14 of the returns in the ACE High Yield Total Return Index, suppose we want to test whether there is a difference between the mean daily returns in Period 1 and in Period 2, which is shown in table below, using a 5% level of significance.

	Period 1	Period 2
Mean	0.01775%	0.01134%
Standard deviation	0.31580%	0.38760%
Sample size	445 days	859 days

Answer:

Step 1: State the hypotheses

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 \neq 0$$

Step 2: Identify the appropriate statistic

Here, we use t-distributed test statistic to test whether there is a difference between the mean daily returns in Period 1 and in Period 2 with degrees of freedom = $n_1 + n_2 - 2 = 445 + 859 - 2 = 1,302$.

Step 3: Specify the level of significance

$$\alpha = 5\% \text{ (two tail)}$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.h]: Hypothesis test concerning the equality of the population means of two at least approximately normally distributed populations based on independent random samples with equal assumed variances.

Answer:

Step 4: State the decision rule

- With $df = 1,302$ we use t-table for two-tailed test with significance level of 5%

The t-table

df	0.1	0.05	0.025	0.01	0.005
120	1.289	1.658	1.980	2.358	2.617
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.960	2.326	2.576

- With $\alpha = 0.05$ and a two-tailed test, probability in each tail would be $\alpha/2 = 0.025$, giving $t_{0.025}^{1,303} \approx 1.960 \rightarrow$ We have t-critical = ± 1.960
 \rightarrow We reject H_0 when Test statistic $> +1.960$ or Test statistic < -1.960 .

Step 5: Calculate the test statistic

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(445 - 1) \times 0.31580^2 + (859 - 1) \times 0.38760^2}{445 + 859 - 2} = 0.133$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}} = \frac{(0.01775 - 0.01134) - 0}{\sqrt{\left(\frac{0.133}{445} + \frac{0.133}{859}\right)}} = 0.3009$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.h]: Hypothesis test concerning the equality of the population means of two at least approximately normally distributed populations based on independent random samples with equal assumed variances.

Answer:

Step 6: Make a decision

$t\text{-statistic} = 0.3009 < t\text{-critical} = +1.960$

→ Do not reject H_0

→ The mean daily returns are **not** different for the two time periods at 5% level of significance.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.i]: Hypothesis test concerning the mean difference of two normally distributed populations (Paired comparison test)

Hypothesis test concerning the mean difference of two dependent samples

Purpose:

- This test (also called paired comparison test) is conducted when our samples may be *dependent* and the observations in the two samples both depend on some other factor.
- The purpose of this test is to check whether the means of the differences between observations for the two samples are different.

Requirement: Populations must be normally distributed.

Step 1: State the hypotheses

- $H_0: \mu_d = \mu_{dz}$ versus $H_a: \mu_d \neq \mu_{dz}$
- $H_0: \mu_d \geq \mu_{dz}$ versus $H_a: \mu_d < \mu_{dz}$
- $H_0: \mu_d \leq \mu_{dz}$ versus $H_a: \mu_d > \mu_{dz}$



Step 2: Identify the appropriate test statistic

This test statistic follows t-distribution, which is referred to t-test.

$$t = \frac{\bar{d} - \mu_{dz}}{s_{\bar{d}}} \text{ with } df = n - 1$$

n : the number of paired observations

μ_{dz} : hypothesized mean of paired differences

$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$: sample mean difference

$s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$: standard error of the mean difference

$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$: sample standard deviation

d_i : difference between the i^{th} pair of observations

MODULE 6: HYPOTHESIS TESTING

[LOS 6.i]: Hypothesis test concerning the mean difference of two normally distributed populations (Paired comparison test)

Hypothesis test concerning the mean difference of two dependent samples

Step 3: Specify the level of significance

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.



Step 4: State the decision rule

- We use t-table with consistent significance level to determine t-critical value.
- Reject H_0 if: t-statistic > **upper t-critical** or t-statistic < **lower t-critical**



Step 5: Test statistic

As we state in Step 2, t-statistic is calculated as:

$$t = \frac{\bar{d} - \mu_{dz}}{s_{\bar{d}}}$$



Step 6: Make a decision

We compare calculated Test statistic with critical value and draw appropriate conclusions.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.i]: Hypothesis test concerning the mean difference of two normally distributed populations (Paired comparison test)

Example 14:

Suppose we want to compare the returns of the ACE High Yield Index with those of the ACE BBB Index. We collect data over 1,304 days for both indexes and calculate the means and standard deviations as shown in table below. Using a 5% level of significance, determine whether the mean of the differences is different from zero.

	ACE index (%)	ACE BBB index (%)	Difference (%)
Mean return	0.0157	0.0135	-0.0021
Standard deviation	0.3157	0.3645	0.3622

Answer:

Step 1: State the hypotheses

$$H_0: \mu_d = 0$$

$$H_a: \mu_d \neq 0$$

Step 2: Identify the appropriate statistic

Here, we use t-distributed test statistic to test whether there is a difference between two dependent population mean with degrees of freedom = $n - 1 = 1,304 - 1 = 1,303$.

Step 3: Specify the level of significance

$$\alpha = 5\% \text{ (two tail)}$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.i]: Hypothesis test concerning the mean difference of two normally distributed populations (Paired comparison test)

Answer:

Step 4: State the decision rule

- With $df = 1,302$ we use t-table for two-tailed test with significance level of 5%

The t-table

df	0.1	0.05	0.025	0.01	0.005
120	1.289	1.658	1.980	2.358	2.617
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.960	2.326	2.576

- With $\alpha = 0.05$ and a two-tailed test, probability in each tail would be $\alpha/2 = 0.025$, giving $t_{0.025}^{1,303} \approx 1.960 \rightarrow$ We have t-critical = ± 1.960
 \rightarrow We reject H_0 when Test statistic $> +1.960$ or Test statistic < -1.960 .

Step 5: Calculate the test statistic

$$\bar{d} = -0.0021\%$$

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{0.3622}{\sqrt{1,304}} = 0.01003\%$$

$$t = \frac{\bar{d} - \mu_{dz}}{s_{\bar{d}}} = \frac{-0.0021 - 0}{0.01003} = -0.20937$$

Step 6: Make a decision

$$t\text{-statistic} = -0.20937 > t\text{-critical} = -1.960$$

\rightarrow Do not reject H_0

\rightarrow The mean of the differences is **not** different from zero at 5% level of significance.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j1]: Hypothesis test concerning the variance of a normally distributed population (Chi-square test)

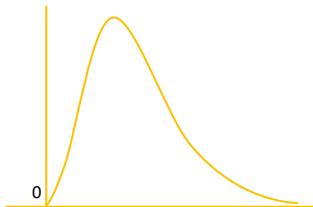
1.

Hypothesis test of single variance (Chi-square test)

Recall the important features of the chi-square distribution

Three important features of the chi-square distribution are:

- It is asymmetrical.
- It is bounded by zero. Chi-square values cannot be negative.
- It approaches the normal distribution in shape as the degrees of freedom increase.



Chi-square test

Purpose: The purpose is to test the value of a *single* population variance.

Requirement: Populations must be normally distributed.

Step 1: State the hypotheses

- $H_0: \sigma^2 = \sigma_0^2$ versus $H_a: \sigma^2 \neq \sigma_0^2$
- $H_0: \sigma^2 \leq \sigma_0^2$ versus $H_a: \sigma^2 > \sigma_0^2$
- $H_0: \sigma^2 \geq \sigma_0^2$ versus $H_a: \sigma^2 < \sigma_0^2$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j1]: Hypothesis test concerning the variance of a normally distributed population (Chi-square test)

1.

Hypothesis test of single variance (Chi-square test)

Chi-square test

**Step 2: Identify
the appropriate
test statistic**

This test statistic follows chi-square distributed, which is referred to as chi-square test.

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

with degrees of freedom = $n - 1$

where:

n is sample size

s^2 is sample variance

σ_0^2 is hypothesized value for the population variance

**Step 3: Specify
the level of
significance**

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j1]: Hypothesis test concerning the variance of a normally distributed population (Chi-square test)

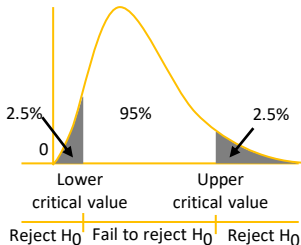
1. Hypothesis test of single variance (Chi-square test)

Chi-square test

Step 4: State the decision rule

- We use chi-square table with consistent significance level to determine chi-square critical value.
- Reject H_0 if:
 test statistic > **upper critical value**
 or test statistic < **lower critical value**

Rejection regions for chi-square distribution for two-tailed test at $\alpha = 5\%$



Step 5: Test statistic

As we state in Step 2, test statistic is calculated as:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Step 6: Make a decision

We compare calculated Test statistic with critical value and draw appropriate conclusions.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j1]: Hypothesis test concerning the variance of a normally distributed population (Chi-square test)

1.

Hypothesis test of single variance (Chi-square test)

Chi-square test

Example 15:

Suppose we are analyzing Sendar Equity Fund, a midcap growth fund that has been in existence for 24 months. During this period, Sendar Equity achieved a mean monthly return of 1.50% and a standard deviation of monthly returns of 3.60%.

Using a 5% level of significance, test whether the standard deviation of returns is less than 4%.

Answer:

Step 1: State the hypotheses

$$H_0: \sigma^2 \geq 16 (\%^2)$$

$$H_a: \sigma^2 < 16 (\%^2)$$

Step 2: Identify the appropriate statistic

Here, we use chi-square distributed test statistic to test whether the standard deviation of returns is less than 4% with degrees of freedom = $n - 1 = 24 - 1 = 23$.

Step 3: Specify the level of significance

$\alpha = 5\%$ (one tail, left side)

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j1]: Hypothesis test concerning the variance of a normally distributed population (Chi-square test)

1.

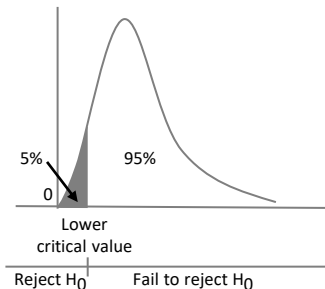
Hypothesis test of single variance (Chi-square test)

Chi-square test

Answer:

Step 4: State the decision rule

- This is one-tailed test and we only focus on the left tail side, so the lower critical value has a 95% probability to its right, which means 5% probability to its left.



MODULE 6: HYPOTHESIS TESTING

[LOS 6.j1]: Hypothesis test concerning the variance of a normally distributed population (Chi-square test)

1.

Hypothesis test of single variance (Chi-square test)

Chi-square test

Answer:

Step 4: State the decision rule

- With $df = 23$ we use chi-square table for one-tailed test with significance level of 5%.
 - The chi-square values in chi-square table correspond to the probabilities in the right tail of the distribution.
- The lower critical value is calculated as the critical value that has a 95% probability to its right (which effectively means 5% to its left).

The chi-square table

df	Probability in right tail								
	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558

→ We have lower critical value = 13.091.

→ We reject H_0 when Test statistic < 13.091.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j1]: Hypothesis test concerning the variance of a normally distributed population (Chi-square test)

1.

Hypothesis test of single variance (Chi-square test)

Chi-square test

Answer:

Step 5: Calculate the test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(24-1) \times (3.60\%)^2}{(4\%)^2} = 18.63$$

Step 6: Make a decision

Test statistic = 18.63 > lower critical value = 13.091

→ Do not reject H_0

→ The standard deviation of returns is greater or equal 4% at 5% level of significance.

MODULE 6: HYPOTHESIS TESTING

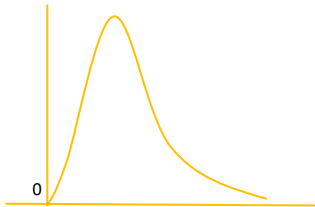
[LOS 6.j2]: Hypothesis test concerning the equality of the variances of two normally distributed populations based on two independent random samples (F-test)

2. Hypothesis test concerning the equality of two variances (F-test)

Recall the important features of the F-distribution

The important features of the F-distribution are:

- It is skewed to the right.
- It is bounded by zero on the left.
- It is defined by two separate degrees of freedom. (df_1 ; df_2)



F-test

Purpose: The purpose is to test the difference between two population variances.

Requirement:

- Populations must be normally distributed.
- The samples are *independent*.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j2]: Hypothesis test concerning the equality of the variances of two normally distributed populations based on two independent random samples (F-test)

2. Hypothesis test concerning the equality of two variances (F-test)

F-test

Step 1: State the hypotheses

- $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_a: \sigma_1^2 \neq \sigma_2^2$
- $H_0: \sigma_1^2 \leq \sigma_2^2$ versus $H_a: \sigma_1^2 > \sigma_2^2$
- $H_0: \sigma_1^2 \geq \sigma_2^2$ versus $H_a: \sigma_1^2 < \sigma_2^2$



Step 2: Identify the appropriate test statistic

This test statistic follows F-distributed, which is referred to as F-test.

$$F = \frac{s_1^2}{s_2^2}$$

$$df_{\text{numerator}} = df_1 = n_1 - 1$$

$$df_{\text{denominator}} = df_2 = n_2 - 1$$

where: s_1^2, s_2^2 are variance of two samples taken from population 1 and population 2.

Note: For easier to determine critical value, we put the larger variance in the numerator when calculating the F-test statistic.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j2]: Hypothesis test concerning the equality of the variances of two normally distributed populations based on two independent random samples (F-test)

2. Hypothesis test concerning the equality of two variances (F-test)

F-test

Step 3:
Specify the
level of
significance

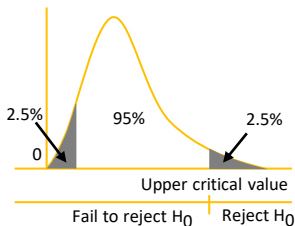
- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.



Step 4: State
the decision
rule

- We use F-table with consistent significance level to determine F-critical value.
 - Since we put the larger variance in the numerator, the F-statistic is always greater than 1 so we consider only the upper critical value.
- Reject H_0 if:
test statistic > **upper critical value**

Rejection regions for F-distribution for two-tailed test at $\alpha = 5\%$



MODULE 6: HYPOTHESIS TESTING

[LOS 6.j2]: Hypothesis test concerning the equality of the variances of two normally distributed populations based on two independent random samples (F-test)

2.

Hypothesis test concerning the equality of two variances (F-test)

F-test

Step 5: Test statistic

As we state in Step 2, F-statistic is calculated as:

$$F = \frac{s_1^2}{s_2^2}$$



Step 6: Make a decision

We compare calculated Test statistic with critical value and draw appropriate conclusions.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j2]: Hypothesis test concerning the equality of the variances of two normally distributed populations based on two independent random samples (F-test)

2. Hypothesis test concerning the equality of two variances (F-test)

Example 16:

Annie Cower is examining the earnings for two different industries. Cower suspects that the variance of earnings in the textile industry is different from the variance of earnings in the paper industry. To confirm this suspicion, Cower has looked at a sample of 31 textile manufacturers and a sample of 41 paper companies. She measured the sample standard deviation of earnings across the textile industry to be \$4.30 and that of the paper industry companies to be \$3.80. Using a 5% significance level, determine if the earnings of the textile industry have a different standard deviation than those of the paper industry.

Answer:

We assume that the sample of 31 textile manufacturers refers to sample 1 and the sample of 41 paper companies refers to sample 2.

Step 1: State the hypotheses

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_a: \sigma_1^2 \neq \sigma_2^2$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j2]: Hypothesis test concerning the equality of the variances of two normally distributed populations based on two independent random samples (F-test)

2. Hypothesis test concerning the equality of two variances (F-test)

Answer:

Step 2: Identify the appropriate statistic

Here, we use F-distributed test statistic to test whether the variance of earnings for companies in the textile industry is equal to the variance of earnings for companies in the paper industry with:

$$df_{\text{numerator}} = df_1 = n_1 - 1 = 31 - 1 = 30$$

$$df_{\text{denominator}} = df_2 = n_2 - 1 = 41 - 1 = 40.$$

Step 3: Specify the level of significance

$$\alpha = 5\% \text{ (two tail)}$$

Step 4: State the decision rule

As F-table only shows the critical value for right-hand tail, we use F-table with significance level of 2.5% to find the upper critical value.

F-distribution table: Critical value for right-hand tail equal to 0.025

$df_2 \backslash df_1$	24	25	30	40	60
30	2.14	2.12	2.07	2.01	1.94
40	2.01	1.99	1.94	1.88	1.80

→ Upper critical value = 1.94

→ With 2 degrees of freedom at level of significant = 5%, we have F-critical value equals 1.94 → We reject H_0 when F-statistic > 1.94.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.j2]: Hypothesis test concerning the equality of the variances of two normally distributed populations based on two independent random samples (F-test)

2. Hypothesis test concerning the equality of two variances (F-test)

Answer:

Step 5: Calculate the test statistic

$$F = \frac{s_1^2}{s_2^2} = \frac{4.3^2}{3.8^2} = 1.28$$

Step 6: Make a decision

F-statistic = 1.2805 < F-critical = 1.94

→ Do not reject H_0

→ The earnings variances of the textile industry are not significantly different from paper industry at 5% level of significance.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.k]: Compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test.

1.

Compare and contrast parametric and nonparametric tests

Parametric tests

A parametric test has at least one of the following two characteristics:

- It is concerned with parameters, or defining features of a distribution.
- It makes a definite set of assumptions.

Nonparametric tests

A non-parametric test is not concerned with a parameter, and makes only a minimal set of assumptions regarding the population.

2.

Situations when to use nonparametric tests

A non-parametric test is used when:

- The data do not meet distributional assumptions.
- There are outliers.
- The data are given in ranks or use an ordinal scale.
- The hypotheses do not concern a parameter.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.I]: 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 of the population correlation coefficient

Purpose:

The purpose of this test is to assess the strength of the linear relationship (correlation) between two variables.

Step 1: State the hypotheses



Step 2: Identify the appropriate test statistic

- $H_0: \rho = 0$ versus $H_a: \rho \neq 0$
- $H_0: \rho \leq 0$ versus $H_a: \rho > 0$
- $H_0: \rho \geq 0$ versus $H_a: \rho < 0$

This test statistic follows a t-distributed, which is referred to as t-test.

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$$

with degrees of freedom = $n - 2$

r is the sample correlation

n is the number of sample observations

MODULE 6: HYPOTHESIS TESTING

[LOS 6.I]: 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 of the population correlation coefficient

Step 3: Specify the level of significance

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.



Step 4: State the decision rule

- We use t-table with consistent significance level to determine t-critical value.
- Reject H_0 if: t-statistic > **upper t-critical** or t-statistic < **lower t-critical**



Step 5: Test statistic

As we state in Step 2, t-statistic is calculated as:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$



Step 6: Make a decision

We compare calculated Test statistic with critical value and draw appropriate conclusions.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.I]: 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 of the population correlation coefficient

Example 17:

An analyst is examining the annual returns for Investment One and Investment Two over 33 years. As the analyst is most interested in quantifying how the returns of these two series are related, so she calculates the correlation coefficient, equal to 0.43051.

Is there a significant positive correlation between these two return series if she uses a 1% level of significance?

Answer:

Step 1: State the hypotheses

$$H_0: \rho \leq 0$$

$$H_a: \rho > 0$$

Step 2: Identify the appropriate statistic

Here, we use t-distributed test statistic to test whether the correlation is significantly different than 0 with degrees of freedom = $n - 2 = 33 - 2 = 31$.

Step 3: Specify the level of significance

$\alpha = 1\%$ (one tail, right side)

Step 4: State the decision rule

- With $df = 31$ we use t-table for one-tailed test with significance level of 1%.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.I]: 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 of the population correlation coefficient

The t-table

df	0.1	0.05	0.025	0.01	0.005
31	1.309	1.696	2.040	2.453	2.744
32	1.309	1.694	2.037	2.449	2.738
33	1.308	1.692	2.035	2.445	2.733

Answer:

- With $\alpha = 0.01$ and a one-tailed test, probability would be $t_{0.01}^{31} = 2.453$
 → We have $t\text{-critical} = +2.453$ → We reject H_0 when $t\text{-statistic} > +2.453$.

Step 5: Calculate the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.43051\sqrt{31}}{\sqrt{1-0.43051^2}} = 2.656$$

Step 6: Make a decision

$$t\text{-statistic} = 2.656 > t\text{-critical} = +2.453$$

→ Reject H_0

→ The population correlation coefficient between the annual returns of these two investments is positive at 1% level of significance.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.I]: 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

Nonparametric test of the population correlation coefficient (Spearman rank correlation test)

Purpose:

- We use nonparametric test of the population correlation coefficient, which is a test based on the **Spearman rank correlation coefficient** when we believe that the population under consideration meaningfully departs from normality.
- Spearman rank correlation test is calculated on the ranks of the two variables within their respective samples.

Step 1: State the hypotheses

- $H_0: \rho = 0$ versus $H_a: \rho \neq 0$
- $H_0: \rho \leq 0$ versus $H_a: \rho > 0$
- $H_0: \rho \geq 0$ versus $H_a: \rho < 0$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.I]: 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

Nonparametric test of the population correlation coefficient (Spearman rank correlation test)

Step 2: Identify the appropriate test statistic

This test statistic follows a t-distributed, which is referred to as t-test.

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

$$\text{with } r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

with degrees of freedom = $n - 2$

n is the number of sample observations

r_s is Spearman rank correlation (all ranks are integer values)

d_i is difference between two ranks



Step 3: Specify the level of significance

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.I]: 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

Nonparametric test of the population correlation coefficient (Spearman rank correlation test)

Step 4: State the decision rule



Step 5: Test statistic



Step 6: Make a decision

- We use t-table with consistent significance level to determine t-critical value.
- Reject H_0 if: t-statistic > **upper t-critical**
or t-statistic < **lower t-critical**

As we state in Step 2, t-statistic is calculated as:

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

We compare calculated Test statistic with critical value and draw appropriate conclusions.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.m]: Explain tests of independence based on contingency table data

Hypothesis test of independence using contingency table data

Purpose: The purpose of this test is to check whether the two characteristics in a contingency table are independent of each other when the data is *discrete* (using a nonparametric test statistic that is chi-square distributed).

Step 1: State the hypotheses



Step 2: Identify the appropriate test statistic

- H_0 : two characteristics are independent.
 - H_A : two characteristics are not independent.
- This is always an **one-tail test** (right side).

This test statistic follows a chi-square distributed, which is referred to as chi-square test.

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

with degrees of freedom = $(r - 1) \times (c - 1)$

$O_{i,j}$ is the number of observations in cell i,j (row i and column j)

$E_{i,j} = \frac{\text{total for row } i \times \text{total for column } j}{\text{total for all columns and rows}}$: is the expected number of observations in cell i,j

r is the number of row categories

c is the number of column categories

i is the index of categories of characteristic 1; $i = 1, 2, \dots$

j is the index of categories of characteristic 2; $j = 1, 2, \dots$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.m]: Explain tests of independence based on contingency table data

Hypothesis test of independence using contingency table data

Step 3: Specify the level of significance

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.

Step 4: State the decision rule

- We use chi-square table for **one-tail test** with consistent significance level to determine chi-square critical value.
- Reject H_0 if: test statistic > **upper critical value**

Step 5: Test statistic

As we state in Step 2, test statistic is calculated as:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Step 6: Make a decision

We compare calculated Test statistic with critical value and draw appropriate conclusions.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.m]: Explain tests of independence based on contingency table data

Hypothesis test of independence using contingency table data

Example 18:

Suppose we observe the following frequency table of 1,594 exchange-traded funds (ETFs) based on two classifications: size (that is, market capitalization) and investment type (value, growth, or blend).

Using a 5% level of significance, test whether the two classifications in a contingency table are independent of each other.

Investment type (i)	Size based on market capitalization (j)			Total
	Small (j = 1)	Medium (j = 2)	Large (j = 3)	
Value (i = 1)	50	110	343	503
Growth (i = 2)	42	122	202	366
Blend (i = 3)	56	149	520	725
Total	148	381	1,065	1,594

Answer:

Before testing the two classifications, we index our three categories of investment type with $i = 1, 2, \text{ or } 3$, and our three categories of size from small to large with $j = 1, 2, \text{ or } 3$. (as illustrated above)

→ Total of row $i = r = 3$

Total of column $j = c = 3$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.m]: Explain tests of independence based on contingency table data

Hypothesis test of independence using contingency table data

Answer:

Step 1: State the hypotheses

H_0 : Size and Investment type are independent

H_a : Size and Investment type are not independent

Step 2: Identify the appropriate statistic

Here, we use chi-square distributed test statistic to test whether the two classifications in a contingency table are independent of each other with degrees of freedom = $(r - 1) \times (c - 1) = (3 - 1) \times (3 - 1) = 4$.

Step 3: Specify the level of significance

$\alpha = 5\%$ (one tail, right side)

Step 4: State the decision rule

- With $df = 4$ we use chi-square table for one-tailed test with significance level of 5%.

The chi-square table

df	Probability in right tail								
	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
3	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750

→ We have upper critical value = 9.488.

→ We reject H_0 when Test statistic > 9.488.

MODULE 6: HYPOTHESIS TESTING

[LOS 6.m]: Explain tests of independence based on contingency table data

Hypothesis test of independence using contingency table data

Answer:

Step 5: Calculate the test statistic

Firstly, we calculate the expected number of observations (expected frequency) in cell i,j using formula $E_{i,j} = \frac{\text{total for row } i \times \text{total for column } j}{\text{total for all columns and rows}}$

Investment type (i)	Size based on market capitalization (j)		
	Small (j = 1)	Medium (j = 2)	Large (j = 3)
Value (i = 1)	$E_{1,1} = \frac{148 \times 503}{1,594} = 46.703$	$E_{1,2} = \frac{381 \times 503}{1,594} = 120.228$	$E_{1,3} = \frac{1,065 \times 503}{1,594} = 336.070$
Growth (i = 2)	$E_{2,1} = \frac{148 \times 366}{1,594} = 33.982$	$E_{2,2} = \frac{148 \times 366}{1,594} = 87.482$	$E_{2,3} = \frac{148 \times 366}{1,594} = 244.836$
Blend (i = 3)	$E_{3,1} = \frac{148 \times 725}{1,594} = 67.315$	$E_{3,2} = \frac{148 \times 725}{1,594} = 173.290$	$E_{3,3} = \frac{148 \times 725}{1,594} = 484.395$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.m]: Explain tests of independence based on contingency table data

Hypothesis test of independence using contingency table data

Answer:

Step 5: Calculate the test statistic

Secondly, we calculate the scaled squared deviation for each combination of size and investment type using formula $\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$

Investment type (i)	Size based on market capitalization (j)		
	Small (j = 1)	Medium (j = 2)	Large (j = 3)
Value (i = 1)	$\frac{(50 - 46.703)^2}{46.703} = 0.233$	$\frac{(110 - 120.228)^2}{120.228} = 0.870$	$\frac{(343 - 336.070)^2}{336.070} = 0.143$
Growth (i = 2)	$\frac{(42 - 33.982)^2}{33.982} = 1.892$	$\frac{(122 - 87.482)^2}{87.482} = 13.620$	$\frac{(202 - 244.836)^2}{244.836} = 7.399$
Blend (i = 3)	$\frac{(56 - 67.315)^2}{67.315} = 1.902$	$\frac{(149 - 173.290)^2}{173.290} = 3.405$	$\frac{(520 - 484.395)^2}{484.395} = 2.617$

$$\rightarrow \chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = 0.233 + 0.870 + 0.143 + 1.892 + 13.620 + 7.399 + 1.902 + 3.405 + 2.617 = 32.081$$

MODULE 6: HYPOTHESIS TESTING

[LOS 6.m]: Explain tests of independence based on contingency table data

Hypothesis test of independence using contingency table data

Answer:

Step 6: Make a decision

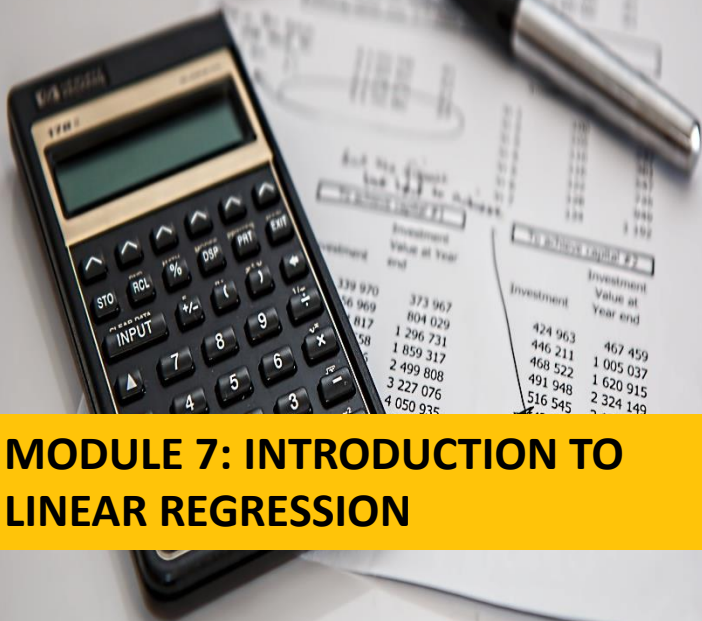
Test statistic = 32.081 > Upper critical value = 9.488

→ Reject H_0

→ The Size and Investment type of ETF are not independent at 5% level of significance.

MODULE 6: HYPOTHESIS TESTING

What we want to test	Test statistic	Probability distribution of the statistic	Degrees of freedom
Test of single mean	$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}; z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	t-distributed z-distributed	$n - 1$
Test of the difference in means	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}}$	t-distributed	$n_1 + n_2 - 2$
Test of the mean of differences	$t = \frac{\bar{d} - \mu_d}{s_d}$	t-distributed	$n - 1$
Test of a single variance	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	Chi-square distributed	$n - 1$
Test of the difference in variances	$F = \frac{s_1^2}{s_2^2}$	F-distributed	$df_1 = n_1 - 1$ $df_2 = n_2 - 1$
Test of a correlation	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	t-distributed	$n - 2$
Test of independence (categorical data)	$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$	Chi-square distributed	$(r-1)(c-1)$



MODULE 7: INTRODUCTION TO LINEAR REGRESSION

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

Learning outcome statement

[LOS 7.a] Describe a simple linear regression model and the roles of the dependent and independent variables in the model.

[LOS 7.b] Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation.

[LOS 7.c] Explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated.

[LOS 7.d] Calculate and interpret the coefficient of determination and the F-statistic in a simple linear regression.

[LOS 7.e] 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.

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance.

[LOS 7.g] 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.

[LOS 7.h] Describe different functional forms of simple linear regressions.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.a] Describe a simple linear regression model and the roles of the dependent and independent variables in the model

The purpose of **simple linear regression** is to explain the variation in a dependent variable in terms of the variation in a single independent variable.

Simple linear model: $Y_i = b_0 + b_1X_i + \epsilon_i$ with $i = 1, \dots, n$
(More detail is presented in next LOS)

The dependent variable (Y)

- The variable whose variation is explained by the independent variable.
- Dependent variable is also referred to as *the explained variable, the endogenous variable, or the predicted variable.*

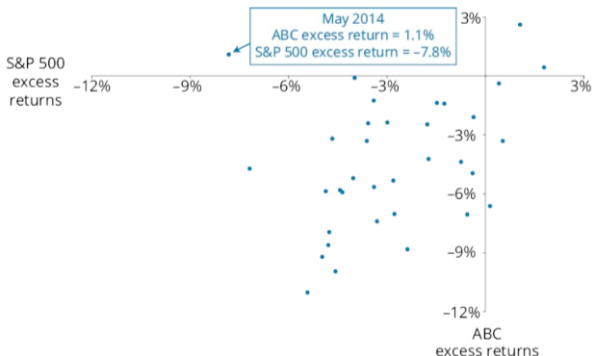
The independent variable (X)

- The variable used to explain the variation of the dependent variable.
- Independent variable is also referred to as *the explanatory variable, the exogenous variable, or the predicting variable.*

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.a] Describe a simple linear regression model and the roles of the dependent and independent variables in the model

Illustration: Use *excess returns on the S&P 500* (independent variable) to explain the variation in *excess returns on ABC Inc. common stock* (dependent variable)



MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.b] Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation

1.

Regression Line

Simple linear regression model

$$Y_i = b_0 + b_1 X_i + \varepsilon_i \text{ with } i = 1, \dots, n$$

Y_i : i^{th} observation of the dependent variable, Y

X_i : i^{th} observation of the independent variable, X

b_0 : The intercept

b_1 : The slope coefficient, which is also known as regression coefficient.

ε_i : The error term, which represents the difference between the observed value of Y and that expected from the true underlying population relation between Y and X .

This simple linear regression relation is described that Y is *regressed* on X

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.b] Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation

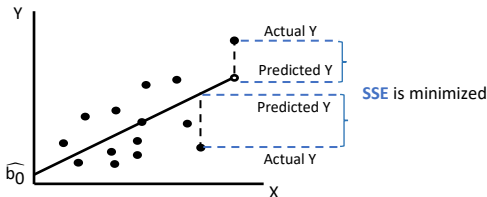
1.

Regression Line

Definition of the Regression Line

- The regression line is the line that **minimizes the sum of the squares error (SSE)** between the Y-values predicted by the regression equation and the actual Y-values.
- It is also the reason why linear regression is also referred to as **ordinary least squares (OLS)** regression.
- The regression line is just one of the many possible lines that can be drawn through the scatter plot of X and Y.

Illustration of the Regression line



MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.b] Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation

1.

Regression Line

Equation of the Regression Line

$$\widehat{Y}_i = \widehat{b}_0 + \widehat{b}_1 X_i \text{ with } i = 1, 2, 3 \dots, n$$

\widehat{Y}_i : Estimated value of Y_i given X_i

\widehat{b}_0 : Estimated intercept term

\widehat{b}_1 : Estimated slope coefficient

Slope coefficient (\widehat{b}_1)

Definition: Describes the change in y for a one unit change in x

Formula: $\widehat{b}_1 = \frac{\text{COV}_{xy}}{\sigma_x^2}$

Intercept term (\widehat{b}_0)

Definition: The line's intersection with the y -axis at $X = 0$

Formula: $\widehat{b}_0 = \bar{Y} - \widehat{b}_1 \bar{X}$
 where: \bar{Y} = mean of Y
 \bar{X} = mean of X

Note: The hat “^” above a variable or parameter indicates a predicted value.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.b] Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation

1.

Regression Line

Interpreting the Regression Coefficients

Intercept term ($\widehat{b_0}$)

The intercept is the value of the dependent variable if the value of the independent variable is zero.

Slope coefficient ($\widehat{b_1}$)

The slope is positive



Change in the independent variable will be in the **same direction** with change in the dependent variable

The slope is negative



Change in the independent variable will be in the **opposite direction** with change in the dependent variable

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.b] Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation

1.

Regression Line

Example 1: Compute the slope coefficient and intercept term for the ABC regression example using the following information.

$$\text{Cov (S\&P 500, ABC)} = 0.000336$$

$$\text{Var (S\&P 500)} = 0.000522$$

$$\overline{\text{S\&P 500}} = -2.70\%$$

$$\overline{\text{ABC}} = -4.05\%$$

Answer:

The slope coefficient is calculated as:

$$\widehat{b}_1 = \frac{\text{cov}_{xy}}{\sigma_x^2} = \frac{\text{Cov (S\&P 500, ABC)}}{\text{Var (S\&P 500)}} = \frac{0.000336}{0.000522} = 0.64.$$

The intercept term is:

$$\widehat{b}_0 = \overline{Y} - \widehat{b}_1 \overline{X} = \overline{\text{ABC}} - \widehat{b}_1 \overline{\text{S\&P 500}} = -4.05\% - 0.64 \times (-2.70\%) = -2.3\%$$

→ Regression line: $\widehat{\text{ABC}} = (-2.3\%) + 0.64 \times \text{S\&P 500}$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.b] Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation

2.

Interpreting the Regression Coefficients

Example 2: Continue with ABC regression example, the estimated slope coefficient was 0.64 and the estimated intercept term was -2.3% . Interpret each coefficient estimate.

Answer:

Regression line: $\widehat{ABC} = (-2.3\%) + 0.64 \times \text{S\&P 500}$

- The intercept term $(\widehat{b_0}) = -2.3\%$

→ *Interpretation:* When the excess return on the S&P 500 is zero, the excess return on ABC stock is -2.3% .

- The slope coefficient $(\widehat{b_1}) = 0.64$

→ Change in the same direction

→ *Interpretation:* When excess S&P 500 returns increase (decrease) by 1%, ABC excess return is expected to increase (decrease) by 0.64%.

Note: The magnitude of the slope coefficient tells us **nothing** about the strength of the linear relationship between the dependent and independent variables.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.b] Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation

3.

Cross-Sectional vs. Time-Series Regressions

Cross-sectional regression

A cross-sectional regression involves many observations of X and Y for **the same time period**. These observations could come from different companies, asset classes, investment funds, countries, or other entities, depending on the regression model.

Example:

Use data from many companies to test whether predicted EPS growth explains differences in price-to-earnings ratios during a specific time period.

$$\widehat{P/E_i} = \widehat{b_0} + \widehat{b_1} \times EPS_i$$

with $i = 1, 2, 3, \dots, n$

Time-series regression

Time-series data use many observations from **different time periods** for the same company, asset class, investment fund, country, or other entity, depending on the regression model.

Example:

Use monthly data from many years to test whether a country's inflation rate determines its short-term interest rates.

$$\widehat{\text{Interest rate}_t} = \widehat{b_0} + \widehat{b_1} \times \text{Inflation}_t$$

with $t = 1, 2, 3, \dots, T$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.c] Explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated

The following four key assumptions are needed to draw valid conclusions from a simple linear regression model:

Linearity

The underlying relationship between the dependent and independent variables is linear.

Homoskedasticity

The variance of the residuals is the same for all observations.

Independence

The observations, pairs of Y_i and X_i , are uncorrelated with one another. This implies the regression residuals are uncorrelated across observations.

Normality

The residuals (ϵ_i), which is how much the observed value of Y_i differs from the \hat{Y}_i estimated using the regression line ($\epsilon_i = Y_i - \hat{Y}_i$), are normally distributed.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.d] Calculate and interpret the coefficient of determination and the F-statistic in a simple linear regression

1. The components of Total sum of squares

In this section, our goal is to explain the variation of the dependent variable. We have: Total variation = Explained variation + Unexplained variation

Total sum of squares (SST) measures the **total variation** in the dependent variable.

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

SST
(Total variation)



RSS
(Explained variation)



SSE
(Unexplained variation)

Regression sum of squares (RSS) measures the variation in the dependent variable that is **explained** by the independent variable.

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

Sum of squared errors (SSE) measures the **unexplained** variation in the dependent variable. It's also known as the sum of squared residuals or the residual sum of squares.

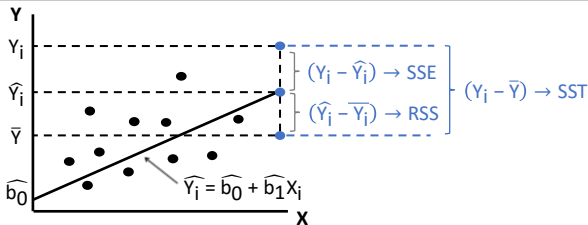
$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.d] Calculate and interpret the coefficient of determination and the F-statistic in a simple linear regression

1. The components of Total sum of squares

Illustration of relationship between SST, RSS and SSE



2. Coefficient of determination (R^2)

The coefficient of determination is the percentage of the variation of the dependent variable that is **explained** by the independent variable.

$$\text{Coefficient of determination } R^2 = \frac{\text{Sum of squares regression (RSS)}}{\text{Sum of squares total (SST)}}$$

→ **Interpretation:** A higher R^2 indicates a better fit for the model.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.d] Calculate and interpret the coefficient of determination and the F-statistic in a simple linear regression

3.

The F - Test

In regression analysis, we use F-distributed test statistic to test whether the slopes are equal to 0, against the alternative hypothesis that at least one slope is not equal to 0.

→ Assesses how well a set of independent variables, **as a group**, explains the variation in the dependent variable.

Step 1: State the hypotheses

Hypothesis test for multiple regression

$H_0 : b_1 = b_2 = b_3 = \dots = b_k = 0$
 H_a : At least one b_k is not equal to 0
where k is the number of independent variables.

Hypothesis test for simple linear regression

$H_0 : b_1 = 0$
 $H_a : b_1 \neq 0$

Note: The F-test in regression analysis is always a one-tailed test.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.d] Calculate and interpret the coefficient of determination and the F-statistic in a simple linear regression

3.

The F - Test

Step 2: Identify the appropriate test statistic

The F – statistic for simple linear regression

For simple linear regression there are only two variances, so $k = 1$:

$$F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/(n-k-1)} = \frac{RSS}{SSE/(n-2)}$$

$df_{\text{numerator}} = k = 1$

$df_{\text{denominator}} = n - k - 1 = n - 2$

Mean square regression (MSR)

$$MSR = \frac{RSS}{k} = \frac{RSS}{1}$$

where:

k = the number of slope parameters estimated (number of independent variables)

Mean square error (MSE)

$$MSE = \frac{SSE}{n-k-1} = \frac{SSE}{n-2}$$

where:

n = number of observations

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.d] Calculate and interpret the coefficient of determination and the F-statistic in a simple linear regression

3.

The F - Test

Step 3: Specify the level of significance

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.

Step 4: State the decision rule

- F-critical depends on the appropriate level of significance and 2 degrees of freedom: $df_{\text{numerator}} = k$ and $df_{\text{denominator}} = n - k - 1$
- And then, we use F-table for **one-tail test** with consistent significance level to determine **F-critical** value.
- Reject H_0 if: F-statistic > **F-critical**

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.d] Calculate and interpret the coefficient of determination and the F-statistic in a simple linear regression

3.

The F - Test

Step 5: Calculate the test statistic

As we state in Step 2, F-statistic is calculated as: $F = \frac{MSR}{MSE} = \frac{RSS}{SSE/(n-2)}$

Step 6: Make a decision

Reject H_0 if: F-statistic > **F-critical**

Do not reject H_0 if: F-statistic \leq **F-critical**

→ **Interpretation:** Rejection of H_0 indicates that the independent variable and the dependent variable have a significant linear relationship.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.e] 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

1.

The standard error of estimate (SEE)

The standard error of estimate measures the distance between the observed values of the dependent variable and those predicted from the estimated regression.

$$SEE = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

→ **Interpretation:** The smaller SEE, the better the fit of the model.

2.

Analysis of variance (ANOVA)

Analysis of variance (ANOVA) is a statistical procedure for analyzing the total variability of the dependent variable.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.e] 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

2.

Analysis of variance (ANOVA)

ANOVA table for Simple linear regression

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F-statistic
Regression (explained)	1	$RSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$MSR = \frac{RSS}{1}$	$F = \frac{MSR}{MSE}$
Error (unexplained)	$n - 2$	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n - 2}$	
Total	$n - 1$	$SST = RSS + SSE$		

ANOVA
table

Standard error of estimate (SEE)

$$SEE = \sqrt{MSE}$$

$$R^2 = \frac{RSS}{SST}$$

Test the goodness of fit of regression model (how well the regression model fits the data)

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.e] 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

Example 3: Complete the ANOVA table for the ABC return regression from S&P 500 return, and calculate the R^2 and the standard error of estimate (SEE).

The total number of observations (n) is 36.

Use the ANOVA table to calculate F-statistic. Test null hypothesis at the 5% significance level that the slope coefficient is equal to 0.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	?	0.00756	?
Error (unexplained)	?	0.04064	?
Total	?	?	
R^2	?		
SEE	?		

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.e] 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

Answer 3.1: Complete the ANOVA Table

With $n = 36$, we complete ANOVA Table for ABC Regression:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	1	$RSS = 0.00756$	$MSR = \frac{RSS}{1} = 0.00756$
Error (unexplained)	$n - 2 = 34$	$SSE = 0.04064$	$MSE = \frac{SSE}{n - 2} = \frac{0.04064}{34} = 0.00119$
Total	$n - 1 = 35$	$SST = 0.00756 + 0.04064 = 0.0482$	
R^2	$R^2 = \frac{RSS}{SST} = \frac{0.00756}{0.0482} = 0.157 \text{ or } 15.7\%$		
SEE	$SEE = \sqrt{MSE} = \sqrt{0.0012} = 0.035$		

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.e] 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

Answer 3.2: Use the ANOVA table to calculate F-statistic

Step 1: State the hypotheses

$$H_0: b_1 = 0$$

$$H_a: b_1 \neq 0$$

Step 2: Identify the appropriate statistic

Here, we use F-distributed test statistic to test how well ABC return explains the variation of S&P 500 return with $df_{\text{numerator}} = 1$ and $df_{\text{denominator}} = 34$

Step 3: Specify the level of significance

$\alpha = 5\%$ (one tail, right side)

Step 4: State the decision rule

- With $df_{\text{numerator}} = df_1 = 1$
 $df_{\text{denominator}} = df_2 = 34$

we use F-table for one-tail test with significance level of 5%

F-distribution table: Critical value for right-hand tail equal to 0.05

$df_1 \backslash df_2$	24	25	30	40	60
1	4.26	4.24	4.17	4.08	4.00
2	3.40	3.39	3.32	3.23	3.15

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.e] 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

Answer 3.2: Use the ANOVA table to calculate F-statistic

Step 4: State the decision rule

- Because in the table, we don't have value of $df_2 = 34$, so we will choose the value which lies between 30 and 40.
 - Subsequently, with 2 degrees of freedom at level of significant = 5%, we have F-critical value approximately 4.1.
- We reject H_0 when Test statistic $> +4.1$.

Step 5: Calculate the test statistic

$$F = \frac{MSR}{MSE} = \frac{0.00756}{0.00119} = 6.353$$

Step 6: Make a decision

$$F\text{-statistic} = 6.353 > F\text{-critical} = 4.1$$

→ Reject H_0

→ The slope coefficient is significantly different than 0 at 5% significance level.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.e] 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

Interpretation:

- The coefficient of determination (R^2) indicates that variation in the independent variable explains 15.7% of the variation in the dependent variable.
 - The F-statistic test confirms that the model's slope coefficient (b_1) is different from 0 at the 5% level of significance.
- The model seems to fit the data reasonably well.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

1.

Hypothesis test of slope coefficients (b_1)

We use a **t-distributed** test statistic to test hypotheses about a slope coefficient (b_1).

Example: Testing whether the population slope is different from a specific value or whether the slope is positive.

Note: This is different from F-distributed test which assess how well a **set** of independent variables, as a **group**, explains the variation in the dependent variable

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

1.

Hypothesis test of slope coefficients (b_1)

Step 1: State the hypotheses

- $H_0: b_1 = B_1$ versus $H_A: b_1 \neq B_1$
- $H_0: b_1 \leq B_1$ versus $H_A: b_1 > B_1$
- $H_0: b_1 \geq B_1$ versus $H_A: b_1 < B_1$



Step 2: Identify the appropriate test statistic

We calculate the test statistic as follows:

$$t = \frac{\widehat{b}_1 - B_1}{S_{\widehat{b}_1}}$$

where:

- \widehat{b}_1 is estimated slope coefficient
- B_1 is hypothesized population slope
- $S_{\widehat{b}_1}$ is standard error of slope coefficient

with degrees of freedom = $n - 2$



Step 3: Specify the level of significance

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

1.

Hypothesis test of slope coefficients (b_1)

Step 4: State the decision rule

- We use t-table with consistent significance level to determine t-critical value.
- Reject H_0 if: t-statistic > **+t-critical** or t-statistic < **-t-critical**



Step 5: Test statistic

As we state in Step 2, t-statistic is calculated as:

$$t = \frac{\widehat{b_1} - B_1}{S_{\widehat{b_1}}}$$



Step 6: Make a decision

We use decision rule: reject H_0 if: t-statistic > **+t-critical** or t-statistic < **-t-critical** and draw appropriate conclusions.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

1.

Hypothesis test of slope coefficients (b_1)

Example 4:

The estimated slope coefficient from the ABC example is 0.64 with a standard error equal to 0.26. Assuming that the sample has 36 observations, determine if the estimated slope coefficient is significantly different than zero at a 5% level of significance.

Answer:

Step 1: State the hypotheses

$$H_0: b_1 = 0$$

$$H_a: b_1 \neq 0$$

Step 2: Identify the appropriate statistic

Here, we use t-distributed test statistic to test whether the slope coefficient is significantly different than 0 with degrees of freedom = $n - 2 = 36 - 2 = 34$

Step 3: Specify the level of significance

$$\alpha = 5\% \text{ (two tail)}$$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

1.

Hypothesis test of slope coefficients (b_1)

Answer:

Step 4: State the decision rule

- With $df = 34$ we use t-table for two-tailed test with significance level of 5%

The t-table

df	0.1	0.05	0.025	0.01	0.005
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724

- With $\alpha = 0.05$ and a two-tailed test, probability in each tail would be $\alpha/2 = 0.025$, giving $t_{0.025}^{34} = 2.032 \rightarrow$ We have t-critical = ± 2.032 .
 \rightarrow We reject H_0 when Test statistic $> +2.032$ or Test statistic < -2.032 .

Step 5: Calculate the test statistic

$$t = \frac{\widehat{b_1} - B_1}{s_{\widehat{b_1}}} = \frac{0.64 - 0}{0.26} = 2.46$$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

1.

Hypothesis test of slope coefficients (b_1)

Answer:

Step 6: Make a decision

$t\text{-statistic} = 2.46 > t\text{-critical} = +2.032$

→ Reject H_0

→ The slope is different from 0, or we can say that ABC return is a significant explanatory variable of S&P 500 return.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

2.

Hypothesis test of pairwise population correlation (ρ)

We also use **t-distributed** test statistic to test whether there is a significant correlation (the linear relationship) between dependent and independent variable in the population.

Step 1: State the hypotheses

- $H_0: \rho = 0$ versus $H_a: \rho \neq 0$
- $H_0: \rho \leq 0$ versus $H_a: \rho > 0$
- $H_0: \rho \geq 0$ versus $H_a: \rho < 0$



Step 2: Identify the appropriate test statistic

We calculate the test statistic as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where:

r is the sample correlation

n is the number of sample observations

with degrees of freedom = $n - 2$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

2.

Hypothesis test of pairwise population correlation (ρ)

Step 3: Specify the level of significance

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.

Step 4: State the decision rule

- We use t-table with consistent significance level to determine t-critical value.
- Reject H_0 if: t-statistic $> +t\text{-critical}$ or t-statistic $< -t\text{-critical}$

Step 5: Test statistic

As we state in Step 2, t-statistic is calculated as:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Step 6: Make a decision

We use decision rule: reject H_0 if: t-statistic $> +t\text{-critical}$ or t-statistic $< -t\text{-critical}$ and draw appropriate conclusions.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

2.

Hypothesis test of pairwise population correlation (ρ)

Example 5:

Continue with the ABC example, the sample correlation (r) is 0.396. Assuming that the sample has 36 observations, determine if the population correlation is significantly different than zero at a 5% level of significance.

Answer:

Step 1: State the hypotheses

$$H_0: \rho = 0$$

$$H_a: \rho \neq 0$$

Step 2: Identify the appropriate statistic

Here, we use t-distributed test statistic to test whether the correlation is significantly different than 0 with degrees of freedom = $n - 2 = 36 - 2 = 34$

Step 3: Specify the level of significance

$$\alpha = 5\% \text{ (two tail)}$$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

2.

Hypothesis test of pairwise population correlation (ρ)

Answer:

Step 4: State the decision rule

- With $df = 34$ we use t-table for two-tailed test with significance level of 5%

The t-table

df	0.1	0.05	0.025	0.01	0.005
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724

- With $\alpha = 0.05$ and a two-tailed test, probability in each tail would be $\alpha/2 = 0.025$, giving $t_{0.025}^{34} = 2.032 \rightarrow$ We have t-critical = ± 2.032 .
 \rightarrow We reject H_0 when Test statistic $> + 2.032$ or Test statistic < -2.032 .

Step 5: Calculate the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.396\sqrt{34}}{\sqrt{1-0.396^2}} = 2.515$$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

2.

Hypothesis test of pairwise population correlation (ρ)

Answer:

Step 6: Make a decision

$t\text{-statistic} = 2.515 > t\text{-critical} = +2.032$

→ Reject H_0

→ The population correlation coefficient between ABC return and S&P 500 return is different from 0 at 5% significance level, or we can say that there is a significant correlation between ABC return and S&P 500 return in the population.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

3.

Hypothesis test of the intercept (b_0)

We also use **t-distributed** test statistic to test whether the population intercept is a specific value.

Step 1: State the hypotheses

- $H_0: b_0 = B_0$ versus $H_a: b_0 \neq B_0$
- $H_0: b_0 \leq B_0$ versus $H_a: b_0 > B_0$
- $H_0: b_0 \geq B_0$ versus $H_a: b_0 < B_0$



Step 2: Identify the appropriate test statistic

We calculate the test statistic as follows:

where:

\widehat{b}_0 is estimated intercept

B_0 is hypothesized population intercept

$S_{\widehat{b}_0}$ is standard error of the intercept

with degrees of freedom = $n - 2$

$$t_{\text{intercept}} = \frac{\widehat{b}_0 - B_0}{S_{\widehat{b}_0}}$$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

3.

Hypothesis test of the intercept (b_0)

Step 3: Specify the level of significance

- $\alpha\%$ level of significant states that there is a $\alpha\%$ probability of rejecting a true null hypothesis.
- Level of significant would be given in each exercises, but most common are 10%, 5% and 1%.



Step 4: State the decision rule

- We use t-table with consistent significance level to determine t-critical value.
- Reject H_0 if: t-statistic $> +t\text{-critical}$ or t-statistic $< -t\text{-critical}$



Step 5: Test statistic

As we state in Step 2, t-statistic is calculated as:

$$t_{\text{intercept}} = \frac{\widehat{b_0} - B_0}{s_{\widehat{b_0}}}$$



Step 6: Make a decision

We use decision rule: reject H_0 if: t-statistic $> +t\text{-critical}$ or t-statistic $< -t\text{-critical}$ and draw appropriate conclusions.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

3.

Hypothesis test of the intercept (b_0)

Example 6:

Continue with the ABC example, the estimated intercept is 4.875(%) with a standard error equal to 0.68. Assuming that the sample has 36 observations, determine if the intercept is greater than 3(%) at a 5% level of significance.

Answer:

Step 1: State the hypotheses

$$H_0: b_0 \leq 3$$

$$H_a: b_0 > 3$$

Step 2: Identify the appropriate statistic

Here, we use t-distributed test statistic to test whether the intercept is significantly different than 0 with degrees of freedom = $n - 2 = 36 - 2 = 34$

Step 3: Specify the level of significance

$\alpha = 5\%$ (one tail, right side)

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

3.

Hypothesis test of the intercept (b_0)**Answer:**Step 4: State the decision rule

- With $df = 34$ we use t-table for one-tailed test with significance level of 5%

The t-table

df	0.1	0.05	0.025	0.01	0.005
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724

- With $\alpha = 0.05$ and a one-tailed test, probability would be: $t_{0.05}^{34} = 1.691$
 → We have $t\text{-critical} = 1.691$ → We reject H_0 when Test statistic > 1.691 .

Step 5: Calculate the test statistic

$$t_{\text{intercept}} = \frac{\widehat{b_0} - B_0}{s_{\widehat{b_0}}} = \frac{4.875(\%) - 3.0(\%)}{0.68} = 2.76$$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.f] Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance

3.

Hypothesis test of the intercept (b_0)

Answer:

Step 6: Make a decision

$t\text{-statistic} = 2.76 > t\text{-critical} = 1.691$

→ Reject H_0

→ The intercept is greater than 3(%).

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.g] 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

1.

Definition of predicted value

Predicted value are values of the dependent variable based on:

- The estimated regression coefficients ($\widehat{b}_1, \widehat{b}_0$).
- A prediction about the value of the independent variable (X_p).

Formula of predicted value for a simple linear regression:

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

where:

\widehat{Y} = predicted value of the dependent variable

X_p = forecasted value of the independent variable

Example 7: Given the ABC regression equation:

$$\widehat{ABC} = -2.3\% + 0.64 \times \text{S\&P 500}$$

Calculate the predicted value of ABC excess returns if forecasted S&P 500 excess returns are 10%.

Answer:

According to regression line, we have $\widehat{b}_0 = -2.3\%$, $\widehat{b}_1 = 0.64$ and $X_p = 10\%$, so the predicted value for ABC excess returns is determined as follows:

$$\widehat{ABC} = \widehat{b}_0 + \widehat{b}_1 X_p = -2.3\% + 0.64 \times 10\% = 4.1\%$$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.g] 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

2.

Creating a Prediction Interval

Illustration:

Continue with the ABC example, we will create a prediction interval around the predicted value of ABC excess returns.

Step 1: Predict the value of Y (\hat{Y})



Step 2: Choose a significance level



Step 3: Determine the critical value (t_c)

As we calculate in the previous slide, the predicted value of ABC excess returns equals 4.1%.

Assume in this example, significance level (α) is 5% and this is two-tail test.

With $\alpha = 0.05$ and a two-tailed test, probability in each tail would be $\alpha/2 = 0.025$, giving $t_{0.025}^{34} = 2.032$
→ We have $t\text{-critical} = \pm 2.032$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.g] 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

2.

Creating a Prediction Interval

Step 4: Compute the standard error of the forecast (s_f)

$$s_f = \text{SEE} \sqrt{1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{(n-1)\sigma_X^2}}$$

SEE: The standard error of the estimate

X_p : The forecasted value of the independent variable

σ_X^2 : Variance of independent variable

Illustration (cont.)

Remind that we already have the value of:

- $\sigma_X^2 = 0.000522$ and $\bar{X} = \text{S\&P 500} = -2.70$ (Refer to example 1).
- $n = 36$ and $\text{SEE} = 0.035$ (Refer to example 3).
- $X_p = 10\%$ (Refer to example 7).
- Standard error of the forecast is calculated as:

$$s_f = 0.035 \sqrt{1 + \frac{1}{36} + \frac{(10 - (-2.70))^2}{35 \times 0.000522}} \approx 3.30$$

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.g] 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

2.

Creating a Prediction Interval

Step 5: Compute prediction interval

We compute the $(1 - \alpha)\%$ prediction interval for the prediction as:

$$\hat{Y} \pm t\text{-critical} \times s_f$$

Illustration (cont.)

The 95% prediction interval then becomes:

$$4.1 \pm 2.032 \times 3.30 = 4.1 \pm 6.71$$

→ The 95% prediction interval ranges from -2.61% to 10.81%.

Interpretation:

Given a forecast value for S&P 500 excess returns of 10%, we can be 95% confident that ABC excess returns will be between -2.61% and 10.81%.

MODULE 7: INTRODUCTION TO LINEAR REGRESSION

[LOS 7.h] Describe different functional forms of simple linear regressions

When assumptions of linear relationship is violated transforming one or both of the variables can produce a linear relationship. The appropriate transformation depends on the relationship between the two variables. One often-used transformation is to take the natural log of one or both of the variables.

Log-lin model

The dependent variable is logarithmic but the independent variable is linear:

$$\text{Ln}Y_i = b_0 + b_1X_i$$

**we would have to transform the R^2 and F-statistic to enable a comparison with a lin-lin model*

Lin-log model

The dependent variable is linear but the independent variable is logarithmic

$$Y_i = b_0 + b_1\text{Ln}X_i$$

Log-log model

Both the dependent variable and the independent variable are linear in their logarithmic forms, is also referred to as the double-log model

$$\text{Ln}Y_i = b_0 + b_1\text{Ln}X_i$$

KNOWLEDGE BASE

Scan QR code

SAPP knowledge base provides learners with a basic understanding for each CFA learning module in Vietnamese before studying new modules with lecturers.



Supplementary CFA materials in Vietnamese
(Composed by SAPP)