2026

FALCON NOTES
FRM PART I

Quantitative Analysis

BOOK 2

Falcon Edufin

Falconedufin.com

TABLE OF CONTENTS

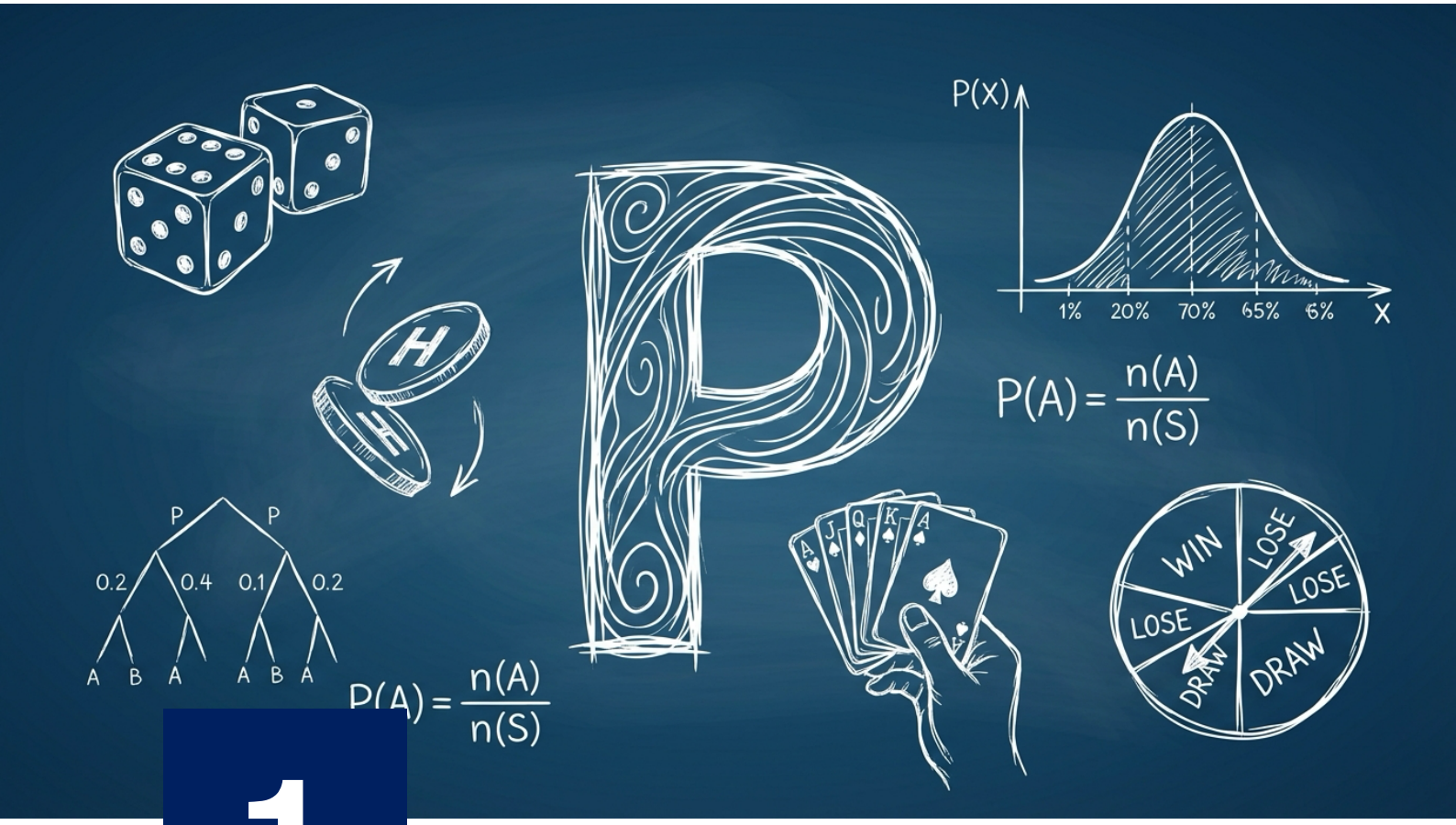
Section 01 Basic Statistics

Section 2 regression

Section 3 time sereis + misc

Table of Contents

Reading 1 Fundamentals of Probabilities



1 Fundamentals of Probability

Scope of this topic

This chapter introduces foundational probability concepts used in risk modeling. It defines events and event spaces, and distinguishes between independent and mutually exclusive events, as well as independent versus conditionally independent events. The chapter develops probability calculations for discrete probability distributions and formalizes conditional probability. It differentiates conditional and unconditional probabilities and applies Bayes' rule to update prior beliefs based on new information.

1.1 Introduction

Probability is a concept well-understood intuitively by most, describing the likelihood of an event's occurrence. It plays a crucial role in decision-making across various aspects of life, from financial investments in stocks or cryptocurrencies to personal health choices like joining a gym or taking yoga classes.

In this section, we will explore key probability concepts essential for risk management, including the classification of events as dependent or independent, the exclusivity of events, and the foundational principles of conditional and unconditional probability. We will also delve into the additive and multiplicative rules governing the probability of dual events and examine Bayes' Theorem for its application in predictive analysis.

1.2 Key Terms

An experiment is a structured process carried out under controlled conditions. If the outcome cannot be predicted in advance, the experiment is referred to as a chance experiment, such as flipping a coin. The result of such an experiment is an outcome, and the set of all possible outcomes is known as the sample space. There are various methods to describe a sample space, which we will address in separate sections:

- Listing all possible outcomes.
- Utilizing a tree diagram.
- Employing a Venn diagram.

The sample space is often denoted by the uppercase letter S or the Greek letter Omega (Ω). For instance, the sample space for flipping a coin once is S = {H, T}, where 'H' represents heads and 'T' represents tails. The representation of the sample space can vary based on the objectives of the experiment. For example, in tossing a coin three times, the sample space could be depicted as the count of heads, S = {0, 1, 2, 3}, or as the sequence of heads and tails, S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}.

An event is a specific set of outcomes from the sample space, and events are typically denoted by uppercase letters such as A or B. For instance, in a single coin flip, event A could represent landing a head. The probability of event A occurring is expressed as P(A).

Theoretical Probability is determined using reasoning or theoretical principles. For a single coin toss, the probability of obtaining heads, based on the logic of two equally likely outcomes, is 50%. This is computed by dividing the number of favorable outcomes (one, head) by the total number of possible outcomes (two, head and tail).

Empirical Probability, on the other hand, is based on actual results from repeated trials. It reflects the long-term frequency of an event's occurrence. For instance, if a six-sided die were rolled 100 times, the empirical probability of each outcome would be calculated from the frequency of that outcome over the 100 rolls.

DIE SIDE	COUNT
1	14
2	20
3	12
4	22
5	25
6	7

In an experiment involving rolling a six-sided die 100 times, the empirical probability of rolling a three, as shown in the table, is 12%, calculated from observing 12 occurrences out of 100 trials. This empirical probability contrasts with the theoretical probability, which is 1/6, or approximately 16.67%, based on the logical premise that each of the six outcomes is equally likely in a fair die roll.

The disparity between empirical and theoretical probabilities stems from the foundational approach each method employs: the former is derived from actual events and trials, while the latter is deduced from established principles assuming perfect conditions.

According to the **Law of Large Numbers,** as the number of trials increases, the empirical probability tends to converge with the theoretical probability. Thus, if the die-rolling experiment were conducted with a larger number of trials, we would expect the empirical probability of rolling a three to approach the theoretical probability of 16.67%. This principle highlights that empirical observations can vary due to chance and that a larger sample size can help achieve results that are more representative of the true probabilities.

1.3 Properties of Probability

Some basic properties of probability.

- The **probability** of any event is always **in between 0 and 1** i.e. For any event, $0 < P(x) < 1$.
- The probability of any event A is equal to the sum of the probabilities of the individual outcomes in A.
- The **sum** of the probabilities of all outcomes in set **must equal 1.** Regardless of whether the set includes equally likely outcomes, this is true.

1.4 Conditional and Unconditional Probabilities

Conditional probability refers to the likelihood of an event occurring given that another event has already happened. For instance, let's consider the probabilities associated with a candidate's progression through financial certifications:

- The probability of passing the FRM exam is denoted as P(A) = 60%.
- Given the candidate passes the FRM exam, the probability of enrolling in the CFA program is P(B | A) = 70%.
- Conversely, if the candidate fails the FRM exam, the probability of then enrolling in the CFA program is P(B | 'A) = 30%.

The conditional probabilities P(B | A) and P(B | 'A) reflect different scenarios contingent upon the initial event of passing or failing the FRM exam, respectively. These are distinct from the unconditional probabilities P(A), P('A), P(B), and P('B), which do not consider the outcome of the FRM exam.

In this context:

- P(B | A) is the probability of enrolling in the CFA program after passing the FRM exam.
- P('B | A) represents the likelihood of not enrolling in the CFA program after passing the FRM exam.
- P(B | 'A) indicates the probability of enrolling in the CFA program after failing the FRM exam.
- P('B | 'A) would be the probability of not enrolling in the CFA program after failing the FRM exam.

The presence of conditional probabilities suggests that the candidate's decision to pursue the CFA program is influenced by the outcome of the FRM exam, thus demonstrating the interconnected nature of these events.

P(A | B) is the formal expression for the conditional probability of A given B.

Formula for calculating conditional probability is.

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

Joint probability is the likelihood of two events happening at the same time. For instance, the joint probability of a candidate passing the FRM exam and enrolling in the CFA program is what we would calculate in our scenario.

It's important not to confuse joint probability with conditional probability. Joint probability, denoted as P(A ∩ B), is the product of the probabilities of both events happening independently, whereas conditional probability, represented as P(B | A), is the probability of one event occurring given that another event has already occurred.

In practice, we can express joint probability in terms of conditional probability:

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

Where Conditional probability is

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

This means you multiply the probability of passing the FRM exam by the probability of enrolling in the CFA program, given that the FRM exam has been passed. This approach is a fundamental concept in statistical analysis and is particularly useful in financial decision-making.

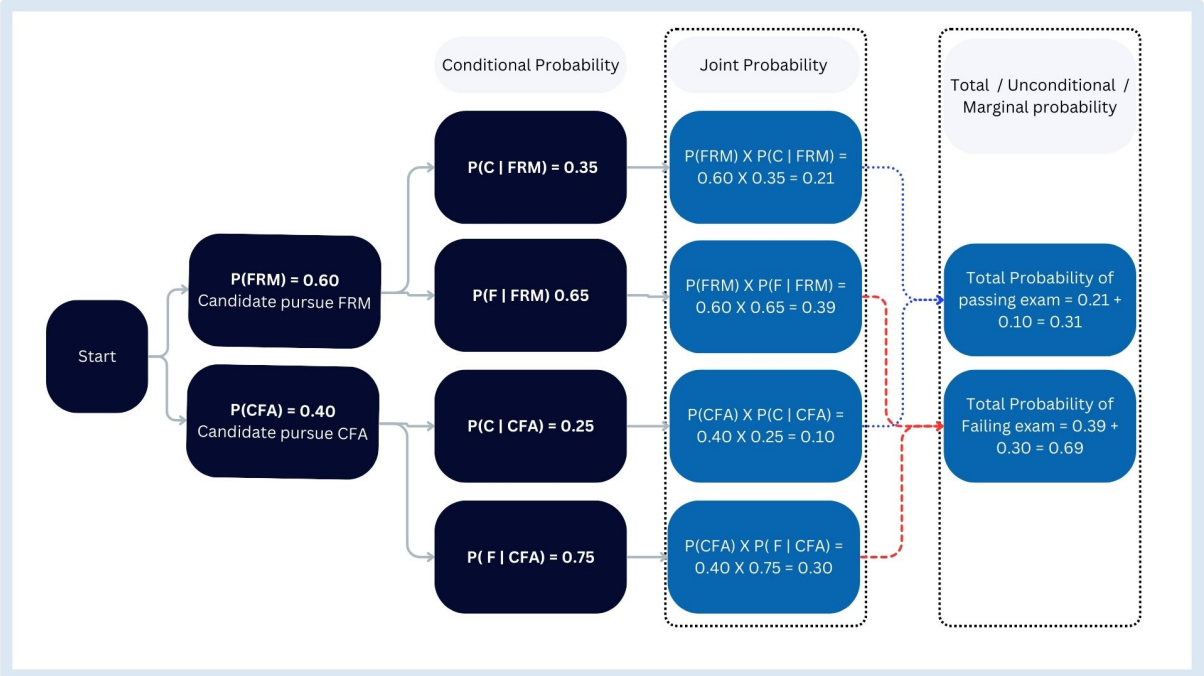
To enhance our understanding of this concept, let's examine the chart presented below.

Chart Explanation (See Diagram below): We will use a different example for this probability tree than the previous one. Here, the candidate has a choice between taking the CFA or FRM exams. The chances of passing either the FRM or CFA exams are given. We should note that in our scenario, choosing the FRM and CFA exams are mutually exclusive events, which means that the candidate can only take one of these exams, not both. Mutually exclusive events, which are also collectively exhaustive (a concept we'll cover in the next section), have probabilities that add up to 100%, as shown in the tree. (It is recommended to look at the tree diagram before reading the total probability explanation.)

Total Probability: In this diagram, we also observe the concept of total probability, sometimes referred to as marginal or unconditional probability. This represents the overall likelihood of either passing or failing an exam, regardless of whether it's the CFA or FRM exam. Total, marginal, or unconditional probability is the aggregate of the joint probabilities of a selected event. In our

example, the probability of passing an exam is calculated by adding the probabilities of choosing to take the FRM and passing, and choosing to take the CFA and passing.

Note: This topic is a popular section that GARP often tests on. You need to be good at using and understanding this probability tree to answer questions in this section. You should be able to go through the tree easily, switching from one part to another.



Note: Following section is part of reading 4 Multivariate random variables, but we prefer to cover it here because this is directly related to conditional probability section.

Probability distribution is the how probability is distributed across the outcomes in sample space. For example, in coin toss experiment, head and tail both has 0.5 probability. We will discuss more about probability distributions in coming readings in detail.

Following table provides the tabular formation of above given chart. This is known as probability matrix. In exam you are more likely to see probability matrix than chart. You should also learn to visualize how chart is created using probability matrix and vice versa. This matrix provides total probabilities and joint probability in intersection point.

	P (CLEARING EXAM)	P(FAILING EXAM)	TOTAL
P(TAKING FRM)	21%	39%	60%
P(TAKING CFA)	10%	30%	40%
TOTAL	31%	69%	100%

Marginal probability distribution: Is the distribution of marginal probabilities for a single variable. Probability mass function (PMF) of exam result is simply 31% + 69% = 100%. GARP might ask you to construct marginal distribution. This is nothing but total probability of a given event, like marginal distribution of course selection is 60% and 40%, which might require calculation of total probability first using the same methods which we learned above.

Conditional Distribution: provides the probabilities of conditional probabilities of each outcome given other specific outcome. This is probability of B given A. Using above example, conditional probability distribution exam results given candidate opted for FRM course = P(C |FRM) and P(F |FRM) = 0.35 + 0.65 = 100%.

1.5 Independent and Mutually exclusive events

The relationship between two events in probability theory is fundamental, as it dictates which rules to apply when calculating probabilities.

Independent events: Two events are independent if the occurrence of one does not affect the probability of the occurrence of the other. For independent events, the probability of both events occurring is the product of their individual probabilities. For instance, if flipping a coin and rolling a die are independent events, the probability of getting a head on the coin and a four on the die is the product of the individual probabilities of these two events.

Mutually exclusive events: Two events are mutually exclusive if they cannot occur at the same time. For mutually exclusive events, the probability of either event occurring is the sum of their individual probabilities. An example of mutually exclusive events is the result of a single card draw from a standard deck: drawing an Ace of Hearts and drawing a King of Spades are mutually exclusive, as one card cannot be both at the same time.

1.5.an Independent events

Two events are considered independent when the occurrence of one event does not influence the probability of the other event occurring. This concept is fundamental in probability theory and has practical implications in various fields, including finance and risk management.

For instance, in the example of Mr. A appearing for both CFA and FRM exams in May 2022, the events of passing or failing one exam do not affect the outcome of the other. Therefore, these two events are independent.

To formally classify two events as independent, they must satisfy specific criteria:

1. Multiplicative Rule: The probability of both events occurring together (joint probability) is equal to the product of their individual probabilities. Mathematically, for two events A and B, they are independent **if $P(A \cap B) = P(A) \times P(B)$.**
2. Consistency in Conditional Probability: The probability of one event occurring given that the other event has occurred is the same as the probability of that event occurring regardless. This means **$P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.**

1.5.b Mutually exclusive events

Mutually exclusive events are those that cannot occur simultaneously. This concept is a key aspect of probability theory and is particularly relevant when assessing scenarios where outcomes are distinctly separate and non-overlapping.

For instance, consider Mr. A's situation of taking the FRM exam in May 2022. The events of passing and failing this exam are mutually exclusive. He can either pass or fail, but both outcomes cannot occur at the same time. The occurrence of one event (passing) automatically excludes the possibility of the other event (failing).

To formalize this concept: Two events A and B are mutually exclusive if the probability of both events occurring together is zero. Mathematically, this is represented as $P(A \cap B) = 0$.

It's important to note that mutually exclusive events are inherently dependent events. The outcome of one event directly influences the probability of the other. In the case of mutually exclusive events, knowing that one event has occurred completely rules out the occurrence of the other.

1.5.c 1.6Addition and Multiplication rule

Events involving connectives "and", "or" and "not":

Example used below: Mr. Mac appeared for both CFA and FRM exam in Nov 2021. Probability of Mac passing FRM exam is P(A) = 0.60 and passing CFA exam is P(B) =0.70.

	Connectives	And	OR
	Written as	P(A and B) P(AB)	P(A or B)
Independent Events	Example	Probability of passing FRM and CFA exam	Probability of passing FRM or CFA exam.
	Rule	Multiplication Rule	Addition rule
	Formula	P(A) X P(B)	P(A)+P(B) – P(AB)
	Solution (assuming independent events)	0.60*0.70 = 0.42	0.60 + 0.70 – 0.42 = 0.88
Mutually exc	Example (modified ignore CFA exam prob)	Probability of passing and failing in FRM exam	Probability of passing or failing FRM exam
	Solution	Both the events cannot happen together hence answer is zero	0.60 + (1-0.60) = 1. *Check note

Note*:- In the sample space of result of FRM exam, there are only two possible outcomes passing and failing. This makes passing and failing event mutually exclusive and collectively exhaustive events. Collectively exhaustive events are all the possible events in event space, which always totals to 1 (like our example).

1.6 Bayes Rule

The fundamental concept of Bayes' Theorem is to update or revise probabilities based on new information. This theorem is particularly useful in scenarios where we have initial assumptions (prior probabilities) and want to update these probabilities upon receiving new evidence or data.

Considering your example, let's say we initially know the probabilities of a student choosing the FRM or CFA course and their respective conditional probabilities of passing. If we later find out that the student has passed the exam (new information), we may want to revise our assessment to determine the probability that the student was enrolled in the FRM course. In probabilistic terms, we are seeking to calculate P(FRM | C), which is the probability of the student having chosen the FRM course given that they have cleared the exam.

Bayes' Theorem provides a formula to calculate this revised probability:

$$P(FRM|C) = \frac{P(C|FRM)P(FRM)}{P(C)}$$

We can also write it as

$$P(FRM \mid C) = \frac{\text{Joint probability of FRM and clearing exam}}{Total\ probablity\ of\ clearing\ exam}$$

Where:

- P(FRM | C) is the probability of the student being in the FRM course given that they have passed the exam.

- P(C | FRM) is the conditional probability of passing the exam given that the student is in the FRM course.

- P(FRM) is the prior probability of a student choosing the FRM course.

- P(C)is the total probability of passing the exam, which can be calculated by considering all relevant pathways or courses that could lead to passing the exam.

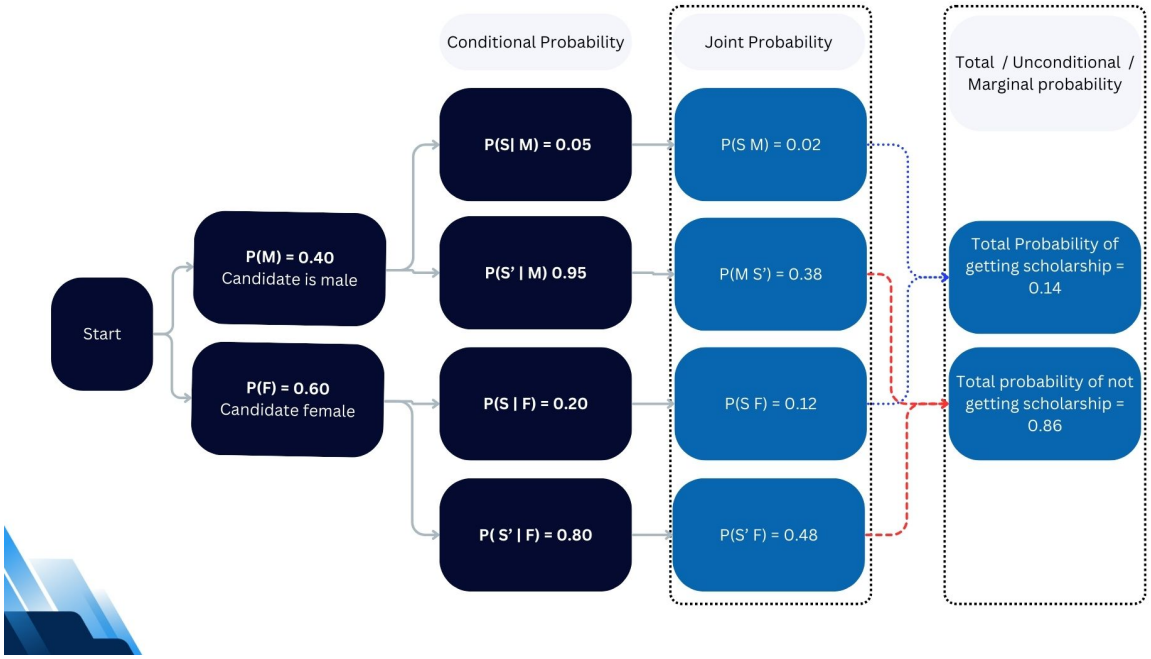
In A and B form we write it as

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

We can similarly find out the probability of student cleared exam and is doing CFA.

Illustration on Bayes:

Falcon University presents a scholarship opportunity for its Risk Management students, requiring candidates to submit an application. Of the 1,000 applications approved for committee review, 400 are from male applicants. The likelihood of a male candidate being awarded a scholarship is 5%, whereas female candidates have a 20% chance. Given these parameters, what is the probability that a scholarship recipient is female?"



Above chart provides the tree diagram for the given case. Where,

- Probability of candidate being male is P(M) and female P(F).
- Probability of scholarship being awarded is P(S).

Question requires you to calculate P(F|S) i.e. probability of candidate being female given scholarship is awarded. To solve this question follow simple recipe –

- Find the joint probability of what is asked and known event. In this case we asked to find the probability of female candidate given scholarship awarded. Hence, we find the joint probability of female candidate and scholarship awarded which goes in the numerator.
- Find the total probability of known event. In this case we know scholarship was awarded, hence the total probability of scholarship awarded goes in the denominator.

$$P(F \mid S) = \frac{Joint\ probability\ of\ female\ and\ scholarship}{Total\ probability\ of\ scholarship} = \frac{C}{A+C} = \frac{C0.12}{0.02+0.12} = 0.857 = 85.7\%.$$

Using the similar method, we can find the following probabilities.

- Probability of candidate is male given the scholarship is awarded = P(M|S) = A / A+C. We can see the denominator is same in above as well as this case. The reason is known event is scholarship is awarded.
- Probability of candidate is male given the scholarship is not awarded = P(M|S') = B / B+D
- Probability of candidate is female given the scholarship is not awarded = P(F|S') = D/ B+D

Exam important note: The illustration provided above is all in one question which covers all the possibilities for the Bayes probability question which is frequently being asked in the exam. But the main challenge in the exam is not the application of the formula but understanding /decoding language of the question.

Note 1: Bayes' area of probability is very vast. FRM exam covers very limited section of this area.

1.7 Conditionally independent events

The concept of conditionally independent events merges two foundational ideas in probability: independent events and conditional probabilities.

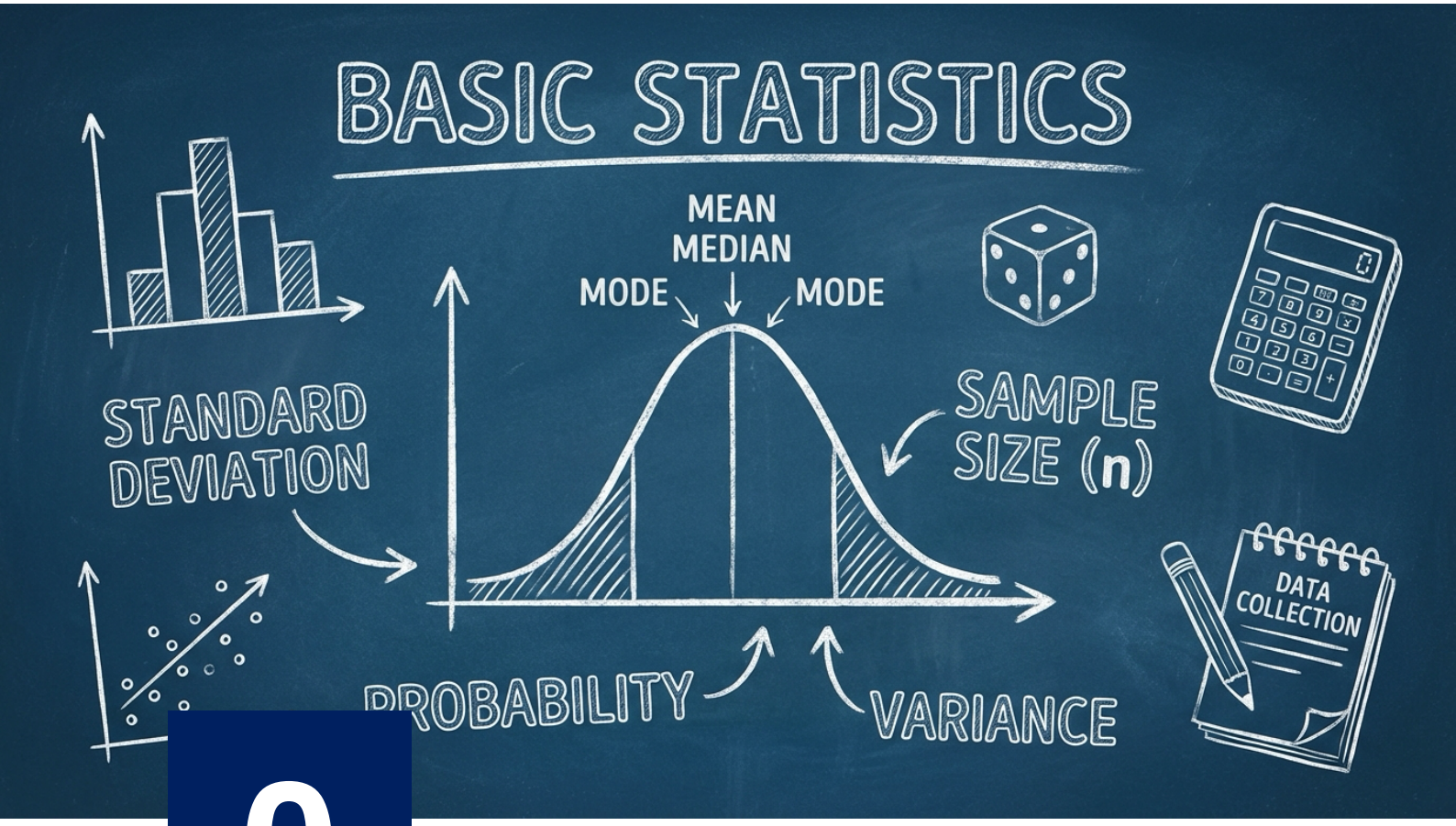
- -Independent Events: We say that two events, A and B, are independent if the probability of both occurring together, $P(A \cap B)$, equals the product of their individual probabilities, $P(A) \times P(B)$.
- -Conditional Independence: When we extend this to conditional events, we consider the probability of A and B given a third event C. Events A and B are conditionally independent given C if the probability of A and B occurring, given C, is the product of their individual probabilities given C. Mathematically, this is expressed as $P(A \mid C) \times P(B \mid C) = P(A \cap B \mid C)$.

It's important to note that the conditional independence of two events given a third event does not necessarily relate to their unconditional independence. This means that two events could be independent under normal circumstances but may become dependent (or vice versa) when considered under a specific condition.

To illustrate, consider the example of Mr. A and Mr. B who work at the same office and live in the same neighborhood. Generally, their arrival times at home are dependent due to shared factors like office hours and transportation. However, if one day Mr. C picks up Mr. A from the office, the likelihood of Mr. A and Mr. B arriving home at the same time becomes independent of each other, as Mr. A's arrival time is now influenced by a different set of circumstances.

This concept, while complex, is crucial in situations where the relationship between events changes under specific conditions, requiring a nuanced understanding of probability and its applications.

Level 0: Basic Statistics (Combination of -Random Variables, Multivariate RV and Sample Moments)



0 Basic Statistics

Scope of this topic

Reading 02: Random Variables
This reading develops the concept of mathematical expectation as the foundation for measuring the central tendency of a random variable. It introduces the four key population moments—mean, variance, skewness, and kurtosis—and explains their interpretation in risk analysis. It also characterizes the quantile function and quantile-based estimators as tools for distributional analysis beyond moments.

Reading 04: Multivariate Random Variables
This reading extends probability concepts to the multivariate setting. It explains how to compute expectations for functions of bivariate discrete random variables and develops covariance and correlation as measures of linear dependence. It analyzes the relationship between covariance, correlation, and independence, examines the impact of linear transformations, and derives the variance of a weighted sum of two random variables.

Reading 05: Sample Moments
This reading focuses on statistical estimation using sample data. It explains how to estimate the mean, variance, standard deviation, and quantiles, including the median, and distinguishes between population and sample moments. It also introduces estimation of joint means and applies the Central Limit Theorem to infer properties of sampling distributions.

L0.1 introduction

'An Apple a day keeps doctor away….'

You work for a research organization and are tasked with verifying the accuracy of this statement. This is a real-world scenario, even if it appears to be completely hypothetical. We keep seeing headlines like this in the media. Following Covid 19, the media has been flooded with headlines such as comorbid people are at higher risk of severe covid infection, XYZ vaccine efficacy is 90%, and so on. These studies are based on statistics. Statistics is the study of data collection and analysis. Statistics is both an art and a science. Science because it is based on well-defined laws and procedures, and art because successful implementation requires human skills and creativity.

We will learn the statistics fundamentals in this chapter. This chapter combines three readings of FRM Part I –Book 2 Quants (the learning objectives are listed above) to cover all basic statistics related concepts at one place. We will stick to the learning objectives outlined in the GARP FRM Part I curriculum.

Returning to our original topic, how can we prove that "an apple a day keeps the doctor away?" Let's look at the statistical procedure step by step. These steps will give you a general idea of how things work in the statistical research field.

Step 1: Comprehending and defining the problem: This is the most critical step where human skills are involved. The problem statement needs human judgment to comprehend and define, and there is no set method to follow. For the rest of the steps, we can rely heavily on established procedures and computer software. In this case, we lack the domain expertise to comprehend and define this statement for finance professionals. This statement would be easier for a nutritionist or doctor to understand and define. If we use our basic knowledge, we can infer, 'eating apple daily keeps body healthy' and therefore less likely to catch infectious diseases. In our professional life, we will deal with problems related to finance and risk management domain.

Step 2: Formulating the hypothesis statement and research planning: A hypothesis is a belief that can be tested statistically. The hypothesis statement derives from the problem statement. We will explore the hypothesis statement and testing more in Reading No 6 Hypothesis Testing. The research plan and design, meanwhile, is the comprehensive approach or method for data collection and analysis. Research can be correlational, descriptive, or experimental, depending on the requirements.

- **Experimental Research**: Statistical research in which two sets are used, one constant and one experimental data. If we consider our example stated above relating to the impact of eating an apple daily on health we will need two sets of candidates, one eating an apple and the other not eating an apple. We will see if people who eat apples daily are less likely to get sick than those who do not. Statements claiming Covid vaccine efficacy are based on experimental research which compares severity of Covid infection in vaccinated and unvaccinated people.
- **Descriptive Research**: Is the study of characteristics of the population. Consider per capita income statistics of Indian population. The average income, median income and standard deviation of income are some of the population parameters

which we need to describe the population income. This type of research is known as descriptive research because it describes the population.

- **Correlational Research**: Is the examination of the relationship between two distinct datasets (variables). As an example, the relationship of levels of vitamin D in blood and morning sun exposure. The more sun ray's exposure one receives, the more Vitamin D in blood. Hence, we can say there is a correlation between Vitamin D and sun exposure.

- **Step 3: Data Collection:** For data research we need to collect the data. Depending on the type of data required for research, data collection can be done through surveys through physical forms or calls. In modern days data collection can be done through online mode such as social networking websites. The collected data can be population data or sample data. The distinction between population and sample is discussed later in this reading.
- **Step 4: Data Summary:** At this stage, we calculate several parameters that serve as a summary of our data. Various parameters including mean, mode, median and variance are used to summarize the data.
- **Step 5: Testing the hypothesis and result interpretation:** The parameters from step 4 are tested using a hypothesis testing approach, and the results are then analyzed.

Note: Only Step 2, 4 and 5 given above are covered in our curriculum.

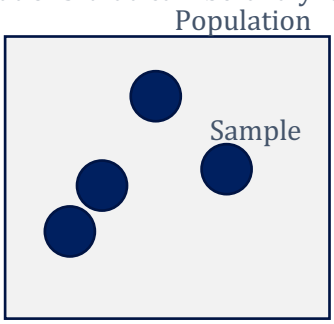
L0.2 Key Terms

Statistics is an art and science of gathering, analyzing, interpreting, and presenting data. There are two domains of statistics.

- Descriptive Statistics is concerned with the organization and summarization of data. Graphs and numerical values are two common ways to summarize data.
- *Inferential Statistics* is a method for drawing conclusions from data. Probability is used in statistical inference to determine how confident we can be that our conclusions are correct.

Data: Facts or figures collected for analysis. Data can be collected from population or sample.

- **Population Data:** Collection of all possible items/observations that can be analyzed. Example, assume you want to analyses the average height of males in India; therefore, all males in India is population data. Gathering this type of data is not only expensive, but also impossible. As a result, we employ sampling methods.



- **Sample data** is the collection of randomly selected items/ observation from the population for statistical analysis. It is a more cost-effective alternative to population data. You can gather data on the heights of 20,000 Indian males, for example, and assume that they represent the population. As a result, research based on a representative sample should yield results that are comparable to population data. To be a representative sample, the sample must contain the characteristics of the population.

Variable, denoted by capital letters such as X and Y, is a characteristic of interest for each item of a population. Variables may be numerical or categorical

- **Numerical variables** take on **numerical values** with equal units such as weight in pounds and time in hours.

- **Categorical variables** place the person or thing into a category. For example male, female, name of the cities.

Random Variable (RV): A random variable in probability statistics has a specific value with a specific probability. It is a variable whose value is determined by uncertain events. For example, let X be the result of a die roll. So, X is a random variable. Its values are 1,2,3,4,5, and 6, with a probability of 1/6.

The term "random" in the context of a random variable does not always imply that the result is completely random, and all possible values are unique or equally likely. It is possible that some values are more likely than others. The word "random" simply means "uncertain." We've Reading No. 2 on Random Variables where we will discuss this concept in more detail. Another important property to remember about random variable is it must take a numeric value. Nonnumeric values such as Gender (male, female) is not a value that can be stored in a random variable. If we want to use categorical values in random variables, use 0 for males and 1 for females, or vice versa. A random variable can be discrete or continuous random variable.

- **Discrete random variables**: If the value is **countable** then it is called as discrete random variable. Example, total number of students in a class.
- **Continuous random variable**: If the values **cannot be counted** then it is called as continuous random variable. Example: Rain drops in rain, stars in galaxy, etc. In some cases, for practical purposes we take it as continuous even if it can be counted e.g. share price (like $60.52).

Univariate Data: Data of only one dimension. Example, say we want to analyze stock performance and we only take stock return of that stock. If we use that stock data with return and trading volume then this data is bivariate data because of two dimensions (return and volume).

Multivariate Data: Data of two or more than two dimensions. In simple terms when we take two or more relatable univariates in data set it becomes multivariate data. For two dimensions we generally use word bivariate. However, in Reading 4 Multivariate Random Variables GARP covered learning objectives mainly of bivariate data analysis. Hence, we will stick to it. Example: Stock returns of two or more stocks or one stock and one index.

L0.3 Location and spread Measures – descriptive statistics

In the previous section we discussed the concept of population data and sample data for statistical analysis. In this section we will discuses the statistical measures and how the population data and sample data affects the calculation of these measures. We will use the same data assuming both population and sample data to understand the impact on measures.

Outlet No	Sale coffee Cups (in 100s)
1	50
2	60
3	70
4	30
5	45
6	45
7	45
8	40
9	60
10	45

1Data set of cup sale

As Population Data: You are working as a data analyst for Starbucks. Starbucks wants you to analyze sales data of their outlets located in New York. There are total 10 outlets and table provides total sales figure in 100's (in number of coffee cups sold). We have taken all observations in the case of population data.

As sample data: Same case as above, except for sample data we assume the sample sales data from all the outlets in US. To save data collection cost you randomly selected 10 outlets and collected sales data. Please note that the data collected is same in both the cases.

We can compute different statistic describing the data. With one variable i.e. univariate data we can compute two forms of statistic. Location measures statistic and spread measures statistic. Location statistic provides the center of the data and spread provides the dispersion in data. We will see both one by one.

L0.3.a Measures of Location – Mean Mode and Median

Location measure also known as measure of central tendency gives the center of the data. The visual analysis of central tendency can be done using histogram as shown in figure. We will discuss three different *measures of central tendency* – **mean, mode and median** (there are other measures of central tendency, but we are limiting our discussion to FRM curriculum).

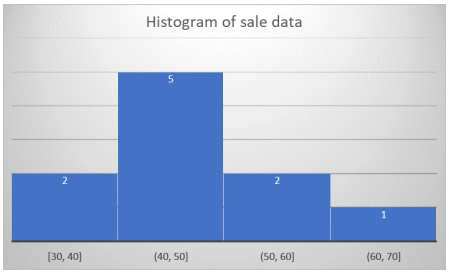


Figure 2 Hist sale data

Mean: also known as arithmetic mean, is most used measure of central tendency. Arithmetic mean is simple average of values.

$$\boldsymbol{Mean} = \frac{\boldsymbol{Sum\ of\ all\ observations}}{\boldsymbol{Number\ of\ observations}} = \frac{\sum x}{n} = \frac{50+60+70+30+45+45+45+40+60+45}{10} = 49$$

Same formula can be written with mathematical notations as

For population mean $\mu = \frac{\sum x}{n}$

For sample mean $\bar{x} = \frac{\sum x}{n}$

$\bar{x}$ (Read as x bar) notation is used for mean when dealing with sample data and μ (read as meu) is used when dealing with population data. $\sum x$ sum of all x and n is total number of observations.

Mode: The value in the data with the highest number of occurrences i.e., has the highest frequency. In some cases, data set may have more than one mode. Such data set is called bimodal for two modes and multimodal for multiple modes. Also, data set may not have any mode if all the values are unique (no value is repeated). In our data set of *total cups sold in outlets*, 45 has the highest frequency of 4.

Outlet No	Sale coffee Cups (in 100s)
4	30
8	40
5	45
6	45
7	45
10	45
1	50
2	60
9	60
3	70

Median: Median is the central value of the data set after sorting. If the count of observations is odd, then median is the observation exactly in the middle. And for even count of observation average of two central value is taken (as there is no central value). In our case we have 10 observations hence average of 5^{th} and 6^{th} observation is used to compute median.

We have 45 in both 5^{th} and 6^{th} position hence the average is 45. Median is 45.

Please note we can calculate mean directly using TI BA II plus calculator however we rarely need to calculate mean in exam. Please refer Falconedufin.com free course on TI BA II Plus calculator.

L0.3.b Measures of Spread: Range, Variance, Standard Deviation

Spread is the measure of dispersion in data. Assume we have two data sets

- Set 1: 12,13 and 14 and
- Set 2: 10, 20 and 30.

We can see values in set 1 are more concentrated whereas values in set 2 are more dispersed. Dispersion is the distance between values. We can check data dispersion using graphical as well as parametric method. The graphical method of data dispersion is discussed in Reading 3 Common univariate random variables. Parametric method of calculating dispersion is covered in this reading. We will study Range, variance -standard deviation and Interquartile range. Each method offers some advantages of dispersion measures which we will discuss one by one. From the risk management perspective, dispersion is the measure of risk. Higher the dispersion in data (return or stock prices) higher the risk. There are other methods of calculating dispersion in data which are not the part of FRM curriculum, hence not discussed here.

Range: Range simplest among all. Range is difference in lowest and highest values of data set. This is very basic information about the data and doesn't offer much value in risk management field. For Set 1 and Set 2 above the range is 2 (14- 12) and 20 (30 – 10) respectively.

Standard Deviation (SD) and Variance: The standard deviation is the most-used measure of dispersion in the field of data analytics, machine learning and risk management. The value of the standard deviation tells how closely the values of a data set are clustered around the mean. In general, a lower value of the standard deviation indicates that the values of that data set are spread over a relatively smaller range around the mean. In contrast, a higher value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean. Variance is calculated in the process of calculating standard deviation and is square of standard deviation. Variance is difficult to interpret in its raw form hence we use square root of variance standard deviation.

$$Vairnace\ of\ population\ =\ \sigma^2 = \frac{\sum(x-u)^2}{n}$$

$$Variance\ of\ sample\ =\ S^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Taking square root of variance will give standard deviation and same can be written in formula.

$$Standard\ deviation\ of\ population\ =\ \sigma = \sqrt{\frac{\sum(x-u)^2}{n}}$$

$$Standard\ Deviation\ of\ sample\ =\ S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Where,

$$\sum(x-u)^2\ is\ sum\ of\ square\ of\ Distance\ to\ mean$$

Why are we dividing sample variance and sample SD by n-1 instead of n like we did in populations case.
Ans: There are multiple answers provided by statisticians. The most prominent one is 'we lose one degree of freedom, hence n-1' which I do not find very convincing. According to me most plausible answer is' when we use samples to estimate SD of population, it is prone to underestimating variance, especially in case of small sample size. Hence reducing 1 from denominator will increase SD. *Example: For sample size of 10 reducing 1 means denominator lowered by 10%, now compare it with sample size of 1000. Reducing 1 from 1000 hardly affects our calculation.*

Note: Don't get intimidated by these formulas. We rarely need to calculate standard deviation using formula in exam. We can use TI BA Plus calculator for calculation of standard deviation (Feed the data and get the answer without using formula).

Steps to calculate Standard Deviation (SD)

Step 1: Calculate mean of values of X. In our case mean is 116.3

Step 2: Calculate distance to mean which is x value – X bar. In the first case it is 135 – 116.3 = 18.70. Repeat this for all values. Sum of distance to mean should always equal to 0 for historical data.

Step 3: Result of the step 2 i.e. sum is zero. This happens due to equal amount of positive and negative distance from the mean. So, our concern is to avoid this sum equal to zero problem. Hence square all the values from the step 2. Squaring will convert -ve sing into positive. And sum squared mean deviation which is 1198.1.

Step 4: Calculation of variance and SD. Once we get sum of squared mean deviation (1198.1) we can now proceed with variance and SD.

	X	X - Xavg	(X - Xavg)^2
Outlet No	Sale coffee Cups (in 100s)	Distance to Mean	Squred Mean Deviation
1	135	18.70	349.69
2	101	-15.30	234.09
3	113	-3.30	10.89
4	131	14.70	216.09
5	111	-5.30	28.09
6	102	-14.30	204.49
7	117	0.70	0.49
8	127	10.70	114.49
9	110	-6.30	39.69
10	116	-0.30	0.09
X bar = avg	116.3	Sum of Squared Distance to Mean	1198.1

VARIANCE OF POPULATION	SD POPULATION	VARIANCE OF SAMPLE	SD OF SAMPLE
1198.1 / 10 =	Root (1198.1/10)	1198.1 / (10-1)	Root (1198.1/(10-1))
119.81	10.9457	113.1222	11.5378

Question 1: How to decide which formula (population or sample) to use in exam for standard deviation calculation?

Ans: GARP will provide this information directly (in major cases) or indirectly in the form of language or case (like analyst selected 30 samples). In case you are not provided with any information (directly or indirectly), use sample data calculations.

Question 2: How to use the calculator to find variance and SD?

Ans: Calculator will only give SD for population and sample both but not variance. To calculate variance (rarely needed) square SD values. We have free course available on TI BA II Plus calculator course which will help you in understanding use of calculator. Simply google TI BA II Plus calculator course by Falcon Edufin.

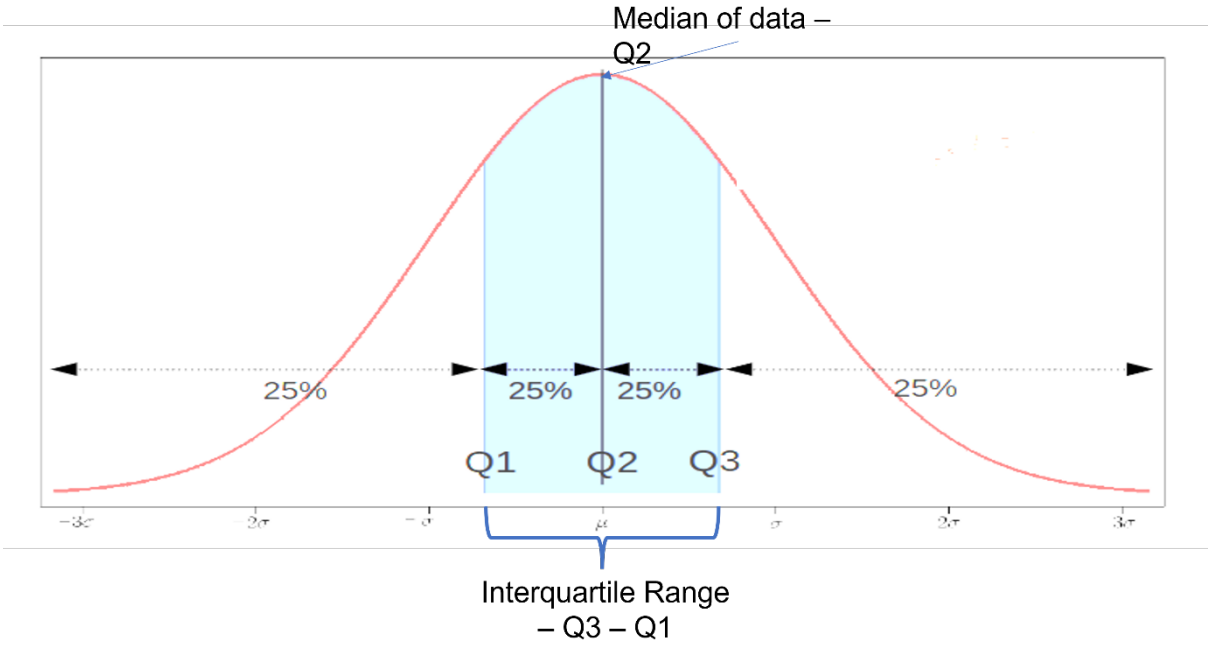
L0.3.c Quantile, quartile, and interquartile range (IQR)

Quartiles are the summary measures that divide a ranked data (after sorting in increasing order) set into four equal parts using three points (check Fig below). These three measures are Q1 (first quartile), Q2 (second quartile), and Q3 (third quartile). Note that Q1 and Q3 are also called the lower and the upper quartiles, respectively.

The second quartile is the same as the median of a data set.

The difference between the third quartile and the first quartile for a data set is called the interquartile range (IQR), which is a measure of dispersion.

Quartile is the part of quantile system which is used to divide data into number of equal parts. Deciles divide data into 10 groups, percentile divides data into 100 groups and quartiles divide data into 4 groups.



Example: You are provided mock exam scores of 100 FRM Part I students of Falcon. We want you to find out the quartiles of exam scores. To find out quartiles first sort the data in ascending order. Following are the quartiles and its interpretation (data and cut-off points are assumed).

PERCENTAGE CUT-OFF	QUARTILE	INTERPRETATION	NOTE
FIRST 25% SCORES CUT-OFF – 42	Q1 – First quartile	Lowest first 25% scorers are equal to or below 42 points	
25 TO 50% SCORES ARE IN BETWEEN 43 TO 56	Q2 – Second quartiles	Second 25% scores lie in between 43 to 56 points	Q2 cut-off is median
50 TO 75% SCORES ARE IN BETWEEN 57-70	Q3 – Third quartile	Third 25% scores lie in between 57 and 70	Q3 Cut-off
ABOVE 70		Top 25% scores are above 70	

The interquartile range is cut-off of 1st and 3rd quartile which is 43 to 70. 2nd quartile the is median.

IQR Vs Standard Deviation

IQR and SD are both the measure of dispersion in data, but these measures are not directly comparable to each other. We can't compare standard deviation of one variable and IQR of another variable and draw conclusion. IQR of one variable is comparable to other and same applicable for SD. Lower IQR and SD means data is more concentrated around the mean. In case of outliers and skewed distribution (will be explained in common univariate topic) IQR is preferred because it is not affected by shape of distribution. Also the change in value of outlier

affects the SD however IQR is not affected by this because it is based on cut off points. IQR remains the same as long as the cutoff points are same.

L0.3.d Expected value and properties of expectation

Till this point we talked about calculation of mean, SD and variance of historical data. What if we want to perform similar set of analysis for future data which is not yet observed like historical data. Let's say you purchased a lottery with 3 winning prizes. $1000, $500, $100 and $50 (In this illustration we are ignoring zero winning situation). What is the average value of prize you can win? To answer this question, we need probability of winning individual prize. Consider the following table providing details of lottery prize and winning probabilities.

With the given information we can calculate the average value of prize using simple method.

Prize in $	Probability
1000	0.10
500	0.20
100	0.30
50	0.40

Avg Prize = 1000 X 0.10 + 500 X 0.20 + 100 X 0.30 + 50 X 0.40

Avg Prize = 250.

This is denoted as E(X) i.e. expected value of X.

Expected Value formula $E(X) = \sum P(x)$

where p is probability of outcome and x is random variable and must total to 1 in every case (0.10+0.20+0.30+0.40)

Hence average prize winning on this lottery is $250. This average is called expected value of a random variable. Expected value is calculated for random variable for given probabilities.

How can someone win $250 in the above example if there is no prize of $250?

Ans: Answer is hidden in true meaning of expected value. Expected value means if we repeat this trail for multiple times then average of all the trails will be equal to expected value i.e. If we buy this lottery a large number of times, our average winnings will be $250.

Calculation of standard deviation of expected value is like what we discussed in previous topic.

$$Variance = \sum\left(\left(x - E(x)\right)^2 * p\right)$$

$$SD = \sqrt{\sum\left(\left(x - E(x)\right)^2 * p\right)}$$

Prize in $	Probability	P* x	X - E(X)	p*(X-E(X)^2)
1000	0.10	100	750	56250
500	0.20	100	250	12500
100	0.30	30	-150	6750
50	0.40	20	-200	16000
	Sum Total E(X)	250	Variance	91500
			SD	302.4896692

Note: We can calculate expected value and SD of expectation using TI BA II Plus calculator. Following table shows the calculation of SD and Variance without calculator just for reference. (Ref Calculator Video)

Expected Value vs Mean

Mean is simple average of the observation with equal weight given. We can calculate mean using similar format of expected value where probability of each observation p = 1/n. 1/n is the equal weight given to each observation. However, in case of **expected value differential weight is given to each value which is probability**.

Useful Properties of Expected Values

Level 0: Basic Statistics (Combination of -Random Variables, Multivariate RV and Sample Moments)

Property 1: E(cX) = c * E(X) where C is constant.

Example: In table LHS consider cX which is 2000 (or any value given below). We can write the same 2000 by separating X = 2 and c=1000 as shown in Table RHS . In RHS C is constant and separated from X.

Table LHS		
cX	Probability	P* cX
1000	0.10	100
2000	0.20	400
5000	0.30	1500
8000	0.40	3200
	E(cX)	5200

Table RHS		
X	Probability	P* X
1	0.10	0.1
2	0.20	0.4
5	0.30	1.5
8	0.40	3.2
	E(X)	5.2
	E(X) * C (1000)	5200

Property 2: E(X+Y) = E(X) + E(Y) if X and Y are independent random variables.

Example:

X	Y	X+Y	p	E(X)	E(Y)	E(X+Y)
100	15	115	0.20	20	3	23
500	30	530	0.35	175	10.5	185.5
800	60	860	0.15	120	9	129
950	80	1030	0.30	285	24	309
				600	46.5	646.5

We can see sum of E(X) 600 and E(y) 46.5 = 645.5 = E(X+Y).

Note: This property is not applicable if X and Y are not independent.

L0.3.e Covariance and Correlation – Multivariate Analysis

Multivariate analysis is the part of separate topic as per FRM curriculum Reading no 4 Multivariate Random Variables some part of which we will cover here itself. When we have two variables for analysis called as bivariate analysis. Measures which deal with bivariate data are covariance, correlation, co-skewness, and co- kurtosis. Correlation is the most used measure to check the relationship between the two variables. In the risk management it is very useful to know relationship between two variables. For example, we want to know the effect on stock price when markets go up or down. In the process of Correlation calculation, we come across Covariance. Covariance is analogous to variance which measures combined variance of two variables. We can also say Variance is covariance of a variable with itself.

$cov\ (x, y) = \frac{\Sigma(x-\overline{x})(y-\overline{y})}{n-1}$ for sample size of n.

Covariance is difficult to interpret using its value. We can only gauge the direction of co-movement of variables using its sign. Positive covariance indicates positive relationship between variables and vice versa. Standardized version of covariance is called correlation and is more interpretable. To get the correlation of two variables, we simply divide their covariance by their respective standard deviations. This specific method of calculating correlation is called Pearson's Correlation. There are other methods to calculate correlation which we will study in last few readings of this subject.

The Pearson's Correlation Coefficient formula

$$\rho(xy) = \frac{Cov\,(xy)}{\sigma_x \sigma_y}$$

Interpreting correlation is very simple. Correlation tells how two variables move together. If increase(decrease) in x causes increase(decrease) in y, then correlation is positive (negative). Correlation ranges from – 1 to +1 for very simple reason. Generally speaking, if movement of x and y is in same direction all the time then correlation is +1 (i.e. 100% of the times movement in same direction). If movement of x and y is in opposite direction all the times, then correlation is -1 (i.e. 100% of the movement is in opposite direction). Say correlation of 0.40, we interpret it as correlation is positive, but movement was not in the same direction every times. Correlation of 0 between two variables indicates no indication of same direction movement.

	X	Y	(X-Xbar)	(Y-Ybar)	(X-Xbar)(Y-Ybar)
	84.00	655.00	1.20	-12.90	-15.48
	61.00	614.00	-21.80	-53.90	1175.02
	89.00	684.00	6.20	16.10	99.82
	91.00	786.00	8.20	118.10	968.42
	88.00	519.00	5.20	-148.90	-774.28
	85.00	750.00	2.20	82.10	180.62
	96.00	703.00	13.20	35.10	463.32
	86.00	728.00	3.20	60.10	192.32
	56.00	600.00	-26.80	-67.90	1819.72
	92.00	640.00	9.20	-27.90	-256.68
Mean	82.80	667.90			**3852.8**
SD	13.34	79.17835		Covariance	428.09

To calculate correlation

= 428.09 / 13.34*79.17

= 0.40

If two variables are highly correlated, it is often the case that one variable causes the other variable, or that both variables share a common underlying driver. Correlation does not provide causation.

Similarly, if two variables are uncorrelated, it does not necessarily follow that they are unrelated. For example, a random variable that is symmetrical around zero and the square of that variable will have zero correlation.

Note 1 (directly testable in exam): Using this formula, if correlation is zero, this does not mean there is no correlation between two random variables. The 0 Pearson's correlation only indicates there is no linear correlation, but variables may have some nonlinear correlation. We have other different methods of testing correlation like Spearman's correlation and Kendal's Tau which are nonlinear correlation measures.

Note 2: This method of correlation calculation is Pearson's correlation coefficient. Later in this subject we will discuss some other measures of correlation.

L0.4 Four common population moments

The population moments used most are which we will cover in this topic are.

- Mean
- Variance
- Skewness
- Kurtosis

We already discussed mean in this reading, which is measure of center of data. Rest of the moments mentioned here are central moments because of measurement uses mean as reference point (X - μ).

Variance is second central moment which measures how data is dispersed as discussed in this reading. Variance is always positive because it is squared term σ^2. Standard deviation (standardized version of variance) is also positive because it is square root of variance σ.

Formula for variance using expectations

$$\sigma^2 = E\{[X - E(X)]^2\} = E[(X - \boldsymbol{\mu})^2]$$

L0.4.a Skewness

Third central moments tell us how symmetrically the data is distributed around the mean. Similar to above equation (used for second central moment), we can calculate third central moment.

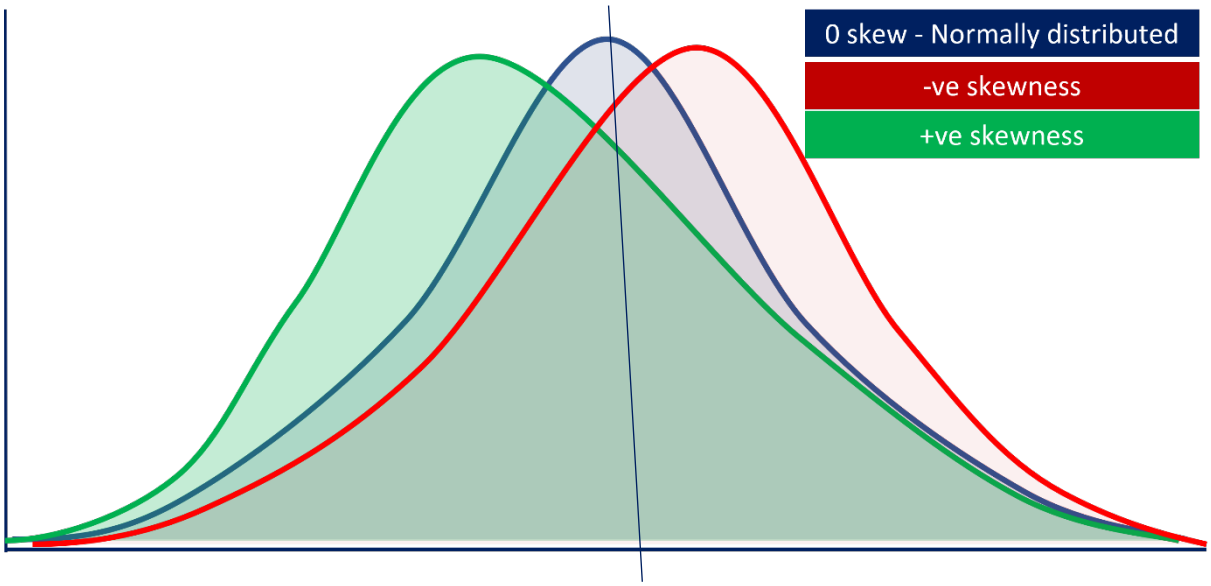
Formula used (not preferred) = σ = $E\{[X - E(X)]^3\} = E[(X - \boldsymbol{\mu})^3]$

Instead of using above formula we prefer standardized version of this moment called skewness.

$$\text{Skewness} = \frac{E[(X-\mu)^3]}{\sigma^3}$$

Skewness provides the information about the outliers in data with its direction. Assume, in FRM exam majority students scored in the range of 45 to 75 with few exceptions of students scoring 100. In this case 100 is outlier and will generate positive skewness in the data. If we assume some students scored 0 or 1 then these values are outliers and will generate negative skewness. These outliers affect mean but do not affect median (mode is central value and hence not affected by extreme values).

Skewness makes comparison between two random variables easier. Skewness is unaffected by constant i.e. skewness of cX = skewness of X, where c is constant multiplier. Skewness of 0 means data is perfectly distributed around the mean.



Skewness-

- Positive skewness is when the outliers are on the right side.
- Negative skewness is when outliers are on the left side.
- Zero skewness is data is symmetrically distributed around the mean.

Skewness plays key role in risk management. Example: if two stock's returns are same in all aspects but one has negative skewness while the other has zero skewness, stock with negative skew return shows the higher probability negative returns are considered riskier.

POSITIVE SKEWNESS	0 SKEWNESS	NEGATIVE SKEWNESS
MODE < MEDIAN < MEAN	Mean = Mode = Median	Mean < median < mode

You must remember above sequence of mean mode median for positive and negative sequence.

L0.4.b Kurtosis

Like second moment, fourth central moment tells us how spread out a random variable is, but by giving more weight on extreme points. Similar to third moment formula we have formula for fourth moment (simply replace all 3 by 4), but not very useful for our exam as well as in real life. We prefer standardized fourth moment called kurtosis.

$$\text{Kurtosis} = \text{K} = \frac{E[(X-\mu)^4]}{\sigma^4}$$

Two assets with same mean, variance and skewness can have different kurtosis. Higher kurtosis indicates more extreme points i.e. higher probability in tail and opposite is true for lower kurtosis. Kurtosis for the normally distributed (normal distribution concept is explained in Reading No 4) data is 3.

Kurtosis can also be measured by its variation called excess kurtosis. Excess kurtosis is K – 3, which is used to relate kurtosis and skewness in line for normal distribution. For normally

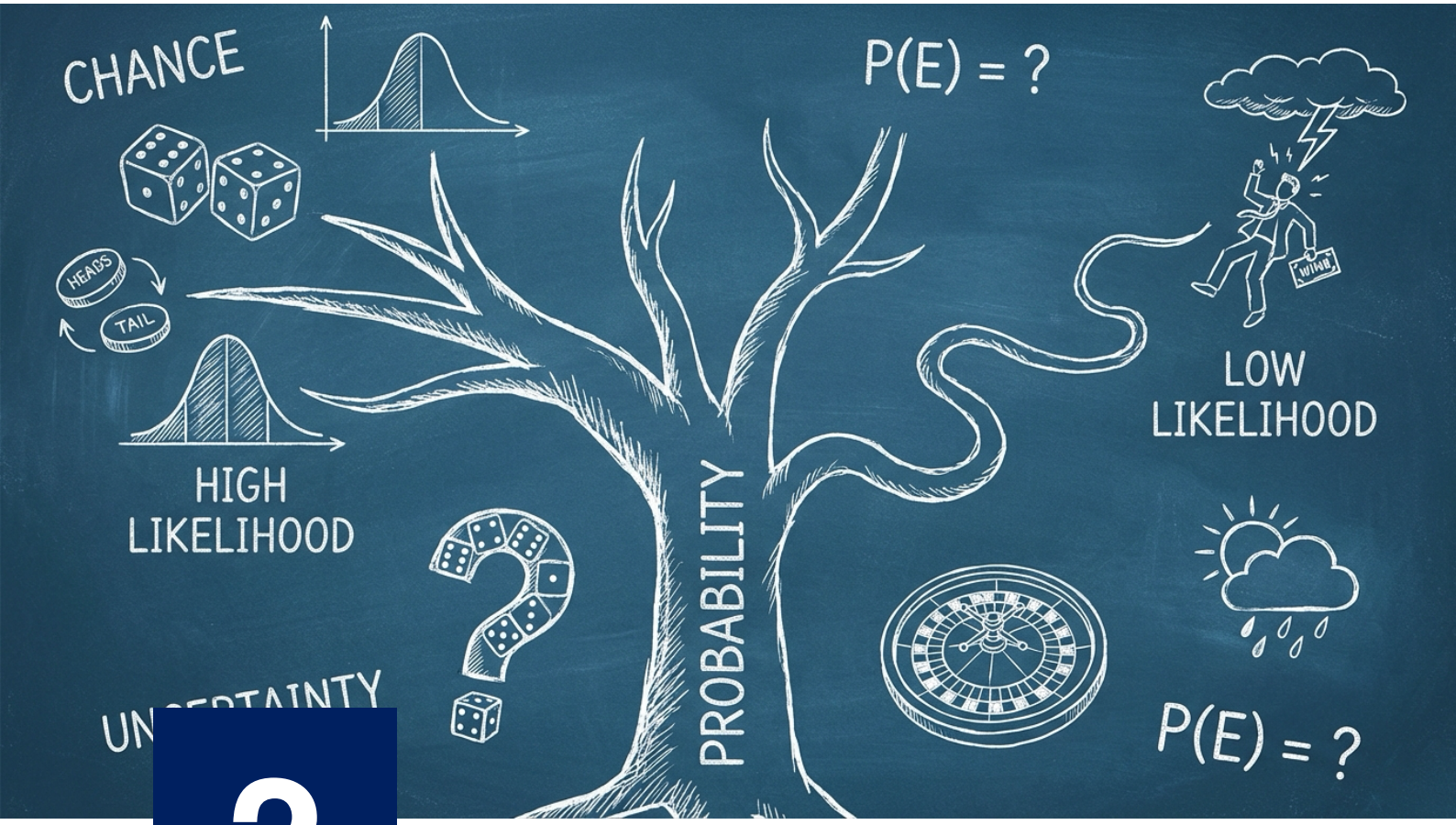
distributed data excess kurtosis is 0 (3 – 3). Distributions with different kurtosis get different names and interpretations as mentioned in this table:

KURTOSIS VALUE	EXCESS KURTOSIS	DISTRIBUTION IS	MEANING
>3	Positive Value	Leptokurtic	Fat tails i.e. More probability in tails and less peaked compared to normal distribution.
=3	Zero	Mesokurtic	Normal distribution
<3	Negative value	Platykurtic	Thin tails i.e. Less probability in tails and more peaked compared to normal distribution.

Note: Mean, Mode and Medians are equal for all the above-mentioned kurtosis.

Note: Questions on skewness and kurtosis calculations are very unlikely in exam, so do not bother about formulas. Focus on the meaning and interpretation of skewness.

Reading 2 Random Variables



2 Random Variables

SCOPE OF THIS TOPIC

This chapter develops the foundational properties of probability distributions. It distinguishes between probability mass functions, probability density functions, and cumulative distribution functions, and explains their interrelationships. The chapter formalizes mathematical expectation and introduces the four key population moments—mean, variance, skewness, and kurtosis—as measures of distributional characteristics. It also characterizes the quantile function and quantile-based estimators, and analyzes the impact of linear transformations on central tendency, dispersion, higher moments, and quantile-based measures such as the median and interquartile range.

Note: Multiple learning objectives from this reading are covered in Level 0 Reading Basic Statistics.

2.1 Discrete random variables – Distribution function

A probability mass function(PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value. Probability that the random variable takes on a specific value is P(X=x). P(X = x) is probability that a random variable X takes the value x (take a note of capital letter X and small letter x used).

Following is the table that shows the values of x in first column, the probability of Xi in second column and cumulative probablity in third column. This table is probability distribution.

The counterpart of PMF is the cumulative distribution function (CDF), which measures the probability of observing a value less than or equal to the input x. (i.e. Pr(X⩽ x). Because the CDF measure's the total probability that X≤ x, it is monotic and increasing in x. CDF is simply sum of PMF till x value. Table provides PMF and and CDF of 6 sided die experiment.

Outcome	PMF	CDF
x	P(x)	P(X<= x)
1	0.1667	0.1667
2	0.1667	0.3333
3	0.1667	0.5000
4	0.1667	0.6667
5	0.1667	0.8333
6	0.1667	1.0000

A discrete probability distribution function has two key characteristics:

1. Each probability is between zero and one
2. The sum of the probabilities is one.
3. The value return from a PMF must be non-negative

To find out the probability of x using CDF we have to simply solve

$P(4) = P(X \leq 4) - P(X \leq 3) = 0.1667$

Note 1: GARP prefers calling probability function of a discrete random variable, probability mass function which is technical term but used less frequently. Majority books written on this topic simply mentions probability function instead of PMF.

2.2 Continuous Random Variable – Distribution Funciton

In contrast to a discrete random variable, a continuous random variable can take on any value within a given range.

Probability Density Function: Continuous random variable uses a probability density function (PDF) in place of the probability mass function. The PDF f(x) returns a non-negative value for any input in the support of X.

Even if the range that the continuous variable occupies is finite, the number of values that it can take is infinite. For this reason, for a continuous variable, the **probability of any specific value occurring is zero.**

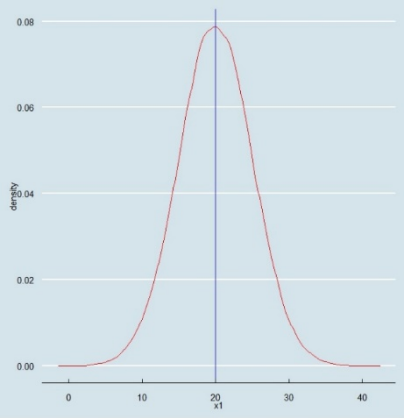
In the case of a continuous random variable, the probability of a specific event happening isn't very clear. But some events are still more likely to happen than others. If we look at 50 years of stock market returns, we might notice that there are more data points between 0% and 5% than between 5% and 10%. In other words, points between 0% and 5% have a lot more of them than the points between 5% and 10% have.

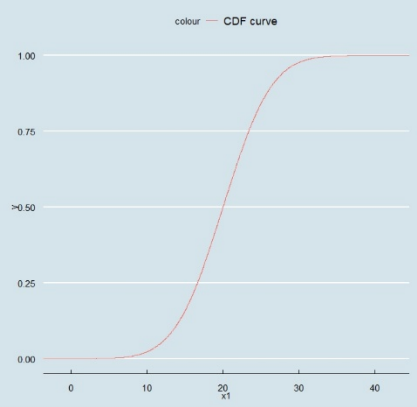
The probability distribution of a continuous random variable possesses the following two characteristics.

- The probability that x assumes a value in any interval lies in the range 0 to 1.
- The total probability of all the (mutually exclusive) intervals within which x can assume a value is 1.0.

Cumulative Distribution Function: Closely related to the concept of a probability density function is the concept of a cumulative distribution function or cumulative density function (both abbreviated CDF). A cumulative distribution function tells us the probability of a random variable being less than a certain value. Traditionally, the cumulative distribution function is denoted by the capital letter of the corresponding density function.

For a random variable X with a probability density function f(x), then, the cumulative distribution function, F(x) is given below in graphical form.





2.3 Linear transformation of random variable

Many variables used in finance and risk management do not have a natural scale. For example, asset returns are commonly expressed as proportions or (if multiplied by 100) as percentages. This difference is an example of a linear transformation. It is helpful to understand the effect of **linear transformations** on the **first four moments** of a random variable.

Let **Y = a + b X**, where **a** and **b** are both **constant** values, it is common to refer to **'a'** as a **location** shift and **'b'** as a **scale**, because these directly affect the mean and standard deviation.

The **mean** of Y is: E(Y) = a + b E(X)

The **variance** of Y is: $b^2V(X) = b^2\sigma^2$

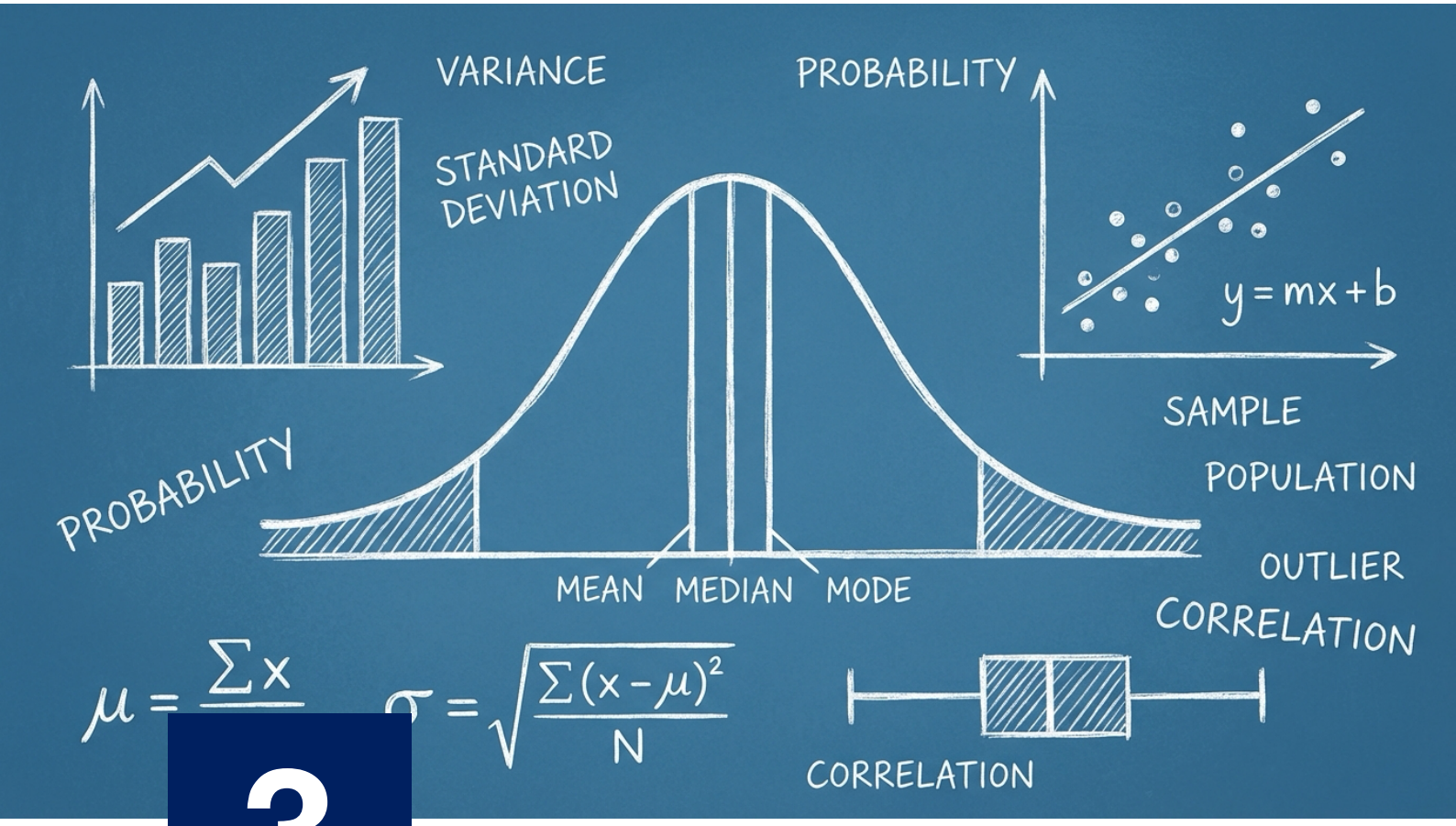
Where, V stands for variance

Note that the location shifts a has no effect on the variance because the variance measures deviation around the mean. The standard deviation of Y is

$$\sqrt{b^2\sigma^2} = |b|\sigma$$

The standard deviation is also insensitive to the shift by a and is linear in b. Finally, if b is positive (so that Y = a + bX is an increasing transformation). Then the skewness and kurtosis of Y are identical to the skewness and kurtosis of X. This is because both moments are defined on standardized quantiles. Which remove effect of the location shift by a and rescaling by b. If b < 0 (and thus Y = a + (-b)X is a decreasing transformation), then the skewness has the same magnitude but the opposite sign. This is because it uses an odd power. The kurtosis which uses an even power (i.e. 4), is unaffected when b<0.

Reading 3 Common Univariate Random Variables



3 Common Univariate Random Variable

Scope of this reading

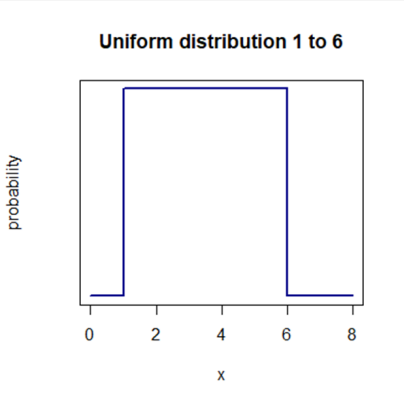
This chapter reviews key probability distributions used in financial modelling. It distinguishes the properties, parameterizations, and typical applications of the uniform, Bernoulli, binomial, Poisson, normal, lognormal, chi-squared, Student's t, and F distributions. The chapter also introduces mixture distributions, explaining how they are constructed by combining component distributions and how they capture features such as skewness, kurtosis, and regime shifts that single standard distributions may not adequately model.

3.1 Introduction

We can start our discussion with the very basic question, what is distribution or probability distribution?

A probability distribution shows the different outcomes that an experiment can have and how likely each one is. For example, the table below shows the probability distribution of rolling a six-sided die. There are two ways to display a distribution, one is graphically, and the other is by using a table. The table format shows the actual distribution. The graphical format is only for illustration and education purposes in real life. In this book, we will use graphs to show different distributions for clarity reasons.

Distribution in Graph Format



Actual PDF and CDF Uniform Distribution

Outcome	PMF	CDF
x	P(x)	P(X<= x)
1	0.1667	0.1667
2	0.1667	0.3333
3	0.1667	0.5000
4	0.1667	0.6667
5	0.1667	0.8333
6	0.1667	1.0000

Distributions can be divided into **two broad categories**: **parametric distributions** and **nonparametric distributions**. A parametric distribution can be described by a mathematical function. In the following sections we explore a few parametric distributions, including the uniform distribution and the normal distribution. A nonparametric distribution cannot be summarized by a mathematical formula. In its simplest form, a nonparametric distribution is just a collection of data. An example of a nonparametric distribution would be a collection of historical returns for a security.

Parametric distributions are often easier to work with, but they force us to make assumptions, which may not be supported by real-world data. Nonparametric distributions can fit the observed data perfectly. The drawback of nonparametric distributions is that they are potentially too specific, which can make it difficult to draw any general conclusions.

Note: For the construction of distribution, mathematical functions are already available. Hence, using distribution is just plug and play for FRM students. We have to only remember formulas for discrete distribution or learn to use readymade tables for continuous distributions for exam purpose.

Illustration No: 3.1

Following table provides total number of smartphones owned by individuals and related probability based on frequency distribution.

Total Number of smartphones owned (X)	Total Individuals (Frequency)	P(X)
0	250	0.0847
1	1600	0.5424
2	800	0.2712
3	300	0.1017
Total	2950	1.0000

Question 1: Find out the probability of a randomly selected individual owns two smartphones.

Solution: P(two smartphones) = P(x) = P(2) = 0.2712 (from the table)

Question 2: Find out the probability of a randomly selected individual owns less than two smartphones.

Solution: Less than two smartphones means either 0 or 1 smartphones owned by individual. We have to apply addition rule here.

3.2 Discrete Distributions

Discrete distribution is the probability distribution of discrete random variable. In our curriculum we have following distributions which we will discuss one by one.

- Discrete Uniform Distribution
- Poisson Distribution
- Binomial Distribution

3.2.a Uniform Distribution

Uniform distribution is the form of distribution where probability is evenly distributed. The uniform distribution can be discrete uniform distribution or continuous uniform distribution depending upon the underlying random variable (discrete or continuous). In this section we will cover both type of distribution to get the better comparison, however continuous uniform distribution belongs to continuous distribution category.

Discrete Uniform Distribution

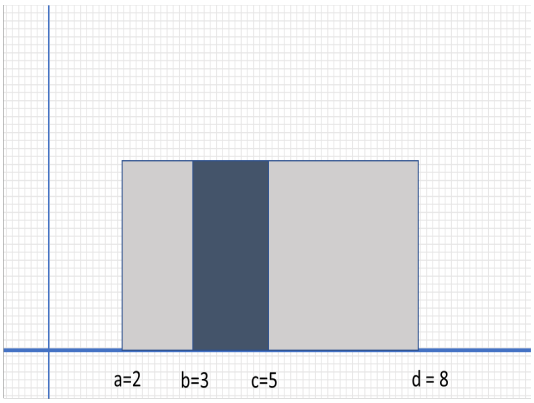
The probability distribution of a discrete random variable lists all the possible values that the random variable can assume and their corresponding probabilities. Six sided die distribution provided above is the example of discrete uniform distribution. Discrete form of distribution is easy to calculate.

$$P(X=x) = \frac{Total\ occurances\ of\ x}{Total\ elements\ in\ sample\ space}$$

In a fair deck of cards, probability of getting queen in randomly drawn card.

$$P(X=queen) = \frac{4}{52} = 7.69\%$$

Continuous Uniform Distribution: Is same as discrete uniform distribution except it supports continuous random variable.



Look at this continuous distribution. We can't calculate the probability of x in continuous random variable. Hence, we calculate the probability of x in specific range P(b<x<c).

$P(b<x<c) = \frac{c-b}{d-a} = 2/6 = 0.333$ or 33.33%.

Properties of continuous uniform distribution for the range a to b (where a is lowest possible value of x and b is the highest value of x) like 2 to 8 in above example,

- Probability density function is $f(x) = \frac{1}{b-a}$
- The mean is $\mu = \frac{a+b}{2}$
- The variance of a uniform distribution, $\sigma^2 = \frac{(b-a)^2}{12}$ (recently tested in exam)

So mean using of above given uniform distribution $= \frac{8+2}{2} = 5$

And variance is $\frac{(8-2)^2}{12} = 36/12 = 3$

3.2.b Bernoulli trails

Bernoulli trail is a random experiment with exactly two possible outcomes, "success" and "failure". Each trail has same probability of success and failure. If probability of success is p then probability of failure is q (1-p). Example of Bernoulli trails are given below,

- Result of Jack in SAT exam (Pass or fail).
- Is the card drawn from deck of card is king of hearts

We will use Bernoulli trials in binomial distribution. In binomial distribution, each trail is Bernoulli trail.

3.2.c Binomial Probability Distribution

Binomial distribution is a special and most widely used discrete probability distribution. It is used to find the probability that an outcome will occur x times in n performances of an experiment. For example, given that 30% of students taking FRM never studied statistics prior to joining FRM, we may want to find the probability that in a random sample of 10 students of FRM, exactly 5 never studied statistics.

There are four conditions that the experiment must meet to be considered a binomial experiment.

Conditions:

1. There are a **fixed number of Bernoulli trials**. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are **only two possible outcomes,** called "success" and "failure," for each trial.
3. The **n trials are independent** and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial.
4. The letter **p denotes the probability of a success** on one trial, and **q denotes the probability of a failure** on one trial, **so p + q = 1**. Since the trials are independent, p stays the same for each trial.

Binomial distribution has two parameters:

- N is the number of independent experiments and
- P is the probability that each experiment is successful

Probability function of binomial distribution

$P(R=r) = nCr \times p^r \times (1-p)^{(n-r)}$

Where, r is value of random variable R. nCr is total r to choose from n trails. (note r and x are same). P is probability of success and 1-p is probability of failure. r or x is total successful trials.

Expected value and variance for x

Illustration No: 3.2

Randomly guessing at a multiple-choice question in FRM exam with 4 possible answers has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose there are 10 multiple choice questions. You guess on each question with no pattern. What is the probability of guessing exactly 6 questions correct?

Solution: To solve any binomial probability question we need n, r and p.

- p is probability of success. For a question with 4 options and 1 option is correct. p of success = ¼ = 0.25
- n is total number of trails. Here we have total 10 questions to try. So n is 10.
- r is total successful trials. Success in our question is correct answer and we want 6 questions correct, hence r = 6.

$P(r=6) = 10C6 \times 0.25^6 (1-0.25)^{(10-6)} = 210 \times 0.00024414 \times 0.3164 = 0.01622$

Hence the probability of getting 6 questions correct using guesswork is 1.622%.

For the given series of n trails,

- Expected value of X = E(X) = n x p
- Variance of X = n x p x (1-p)

Illustration No: 3.3

Assume we have four bonds, each with a 15% probability of defaulting over the next year. The event of default for any given bond is independent of the other bond defaulting. What is the probability of exactly 2 bonds default?

Solution: To solve any binomial probability question we need n, r and p.

What is the mean number of defaults?

The standard deviation?

Solution:

P(R = 2) = $4C2 \times 0.15^2 (1-0.15)^{(4-2)}$

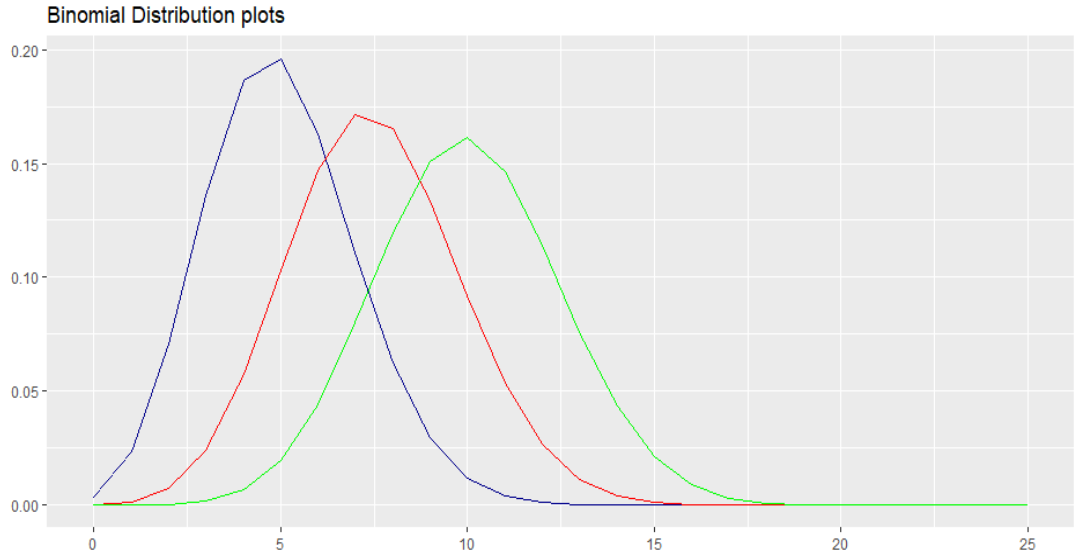
$= 6 \times 0.0225 \times 0.7225$

$= 9.75\%$

$E(X) = 4 \times 0.15 =$

$SD^2 = n\ p\ q = 4 \times 0.15 \times 0.85 = 0.51$

$SD = \sqrt{0.51} = 0.71$



3.2.d Poisson Distribution

Poisson function is used to calculate the probability of specific number of occurrences in given time. Please note, in the Poisson distribution key component is **time**. Following are some of the examples of type of questions you can answer with the help of Poisson distribution.

- What is the probability of raining exactly 60 days in a year.

- What is the probability of 100 customers visiting a mobile shop in a day.
- What is the probability of 3 banks will default within 1 year.

In all the above examples, we have two components, value (x) for which we want to find out the probability, and period. To answer above questions, we need expected value of x (mean of occurrences) for the given time interval. Continuing our example, assume on an average it rains for 80 days in Delhi, India. What is the probability of raining exactly 60 days in a given year? We can use Poisson distribution function to answer this question. Formula for Poisson distribution function is

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Where, λ is mean of occurrences (average of x) and x is value for which we want to find out probability.

Question: Where is the time period in this equation?

Answer: Time period is not considered separately in the equation. If you want to find out x for a year then make sure λ is also for a year. Because both are in same time interval, we do not give separate consideration to time.

Extending our example, assume it rains for 160 days on an average in two years, what is the probability of raining for 60 days in a year. In this extension we can see time interval for λ and x are different. So simply convert λ into one year average and we can fit this into our equation. λ for a year is equal to 160/2 = 80 days. So

$$P(X=60) = \frac{80^{60} e^{-80}}{60!} = 0.003$$

Note: Don't solve this equation in your calculator, this is too heavy for TI Ba II plus calculator. Calculator will show error 1 = overflow value.

Lets take the simple example, A washing machine in a laundromat breaks down an average of three times per month. Using the Poisson probability distribution formula, find the probability that during the next month will have exactly two breakdowns. $P(X=2) = \frac{3^2 e^{-3}}{2!} = 9 \text{ X } (0.04979) / 2 = 0.2240$

Do it yourself

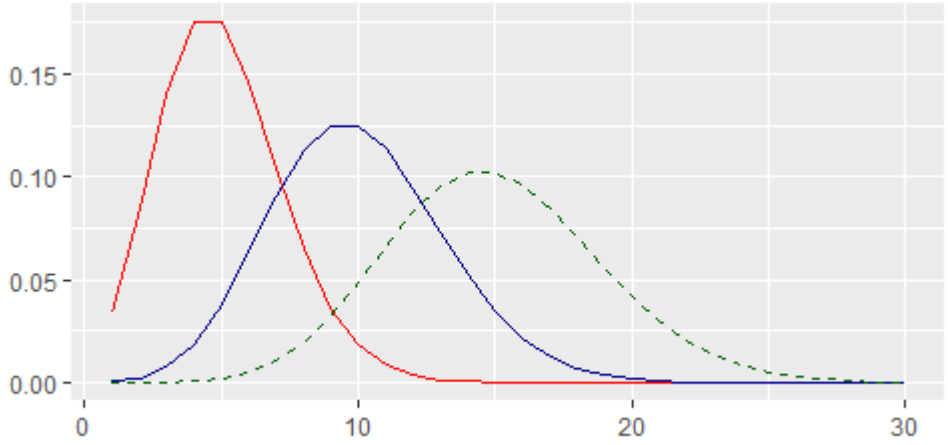
The service desk gets 25 customer complaints in a week of 5 days on average. What is the chance of getting exactly 6 customer complaints in one day? What is the chance of getting fewer than 2 customer complaints in one day?

1. Note 1: Change the time period to 1 day. Average complaints in a day = 25/5 = 5.
2. Note 2: For the second question, we need the addition rule of probability. Fewer than two customers means no customers or one customer calls. Find the probability for each x separately and then add them up.

Poisson distribution is used when events are very rare. This distribution is very important in risk management. Example to find out probability of banks default in a given year, we use Poisson

distribution. Remember Bank's default is very rare event and Poisson distribution helps in finding probability of banks default.

Following is the Poisson distribution for lambda of 5(red), 10(blue) and 15(Dashed) for 1≤ x ≤30 (i.e. various values taken for x ranging 1 to 30).



3.3 Continuous Distributions

We learned in the previous reading Random variable that; continuous random variables are not countable and can assume any value in given interval. This is because infinite number of values are contained in any given interval. Take the example of rate of USD ($) in Indian rupees, $1 = Rs 75.596114. This means there are 1,00,000 possibilities in the USD rate from just Rs 75 to Rs76. Hence probability of any specific value for continuous random variable cannot be found. Assume, in the next month USD to INR rate is likely to move in the range of Rs 72 to Rs 75 per USD. What is the probability of $1 = Rs 75.596400? One can say its 1/ 300000 (possible values in the range). This gives us very small probability. Hence, for continuous random variable, probability is calculated for the range. Assume, $ to ₹ is likely to move in the range of ₹72 to ₹75 in the next month. What is the probability of rate of $ to ₹ between ₹73 to ₹74, assuming probabilities are uniformly distributed. To calculate this we can simply divide 1 interval with the total possible intervals, 3 in this case: ₹72 to ₹73, ₹73 to ₹74 and ₹74 to ₹75. Hence the probability of $ to ₹ will move in the range of ₹73 to ₹74 is 1/3 = 33.33% approx.

Exam Important point: For continuous random variable probability of X = x is always equal to zero.

Probability distribution of continuous random variable are called continuous distribution. Following are the continuous distributions we will cover in our curriculum –

Symmetrical Distributions

- Normal Distribution
- Standard Normal Distribution
- Students t distribution

Nonsymmetrical distribution

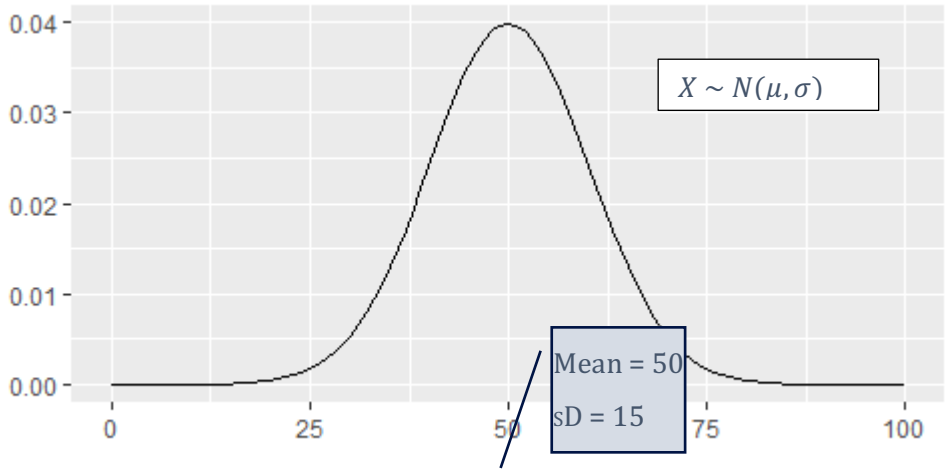
- Lognormal Distribution
- Chi Square Distribution

- F Distribution
- Exponential Distribution
- Beta Distribution

3.3.a Normal Distribution and standard normal distribution

The normal distribution is the most widely used distribution in statistics and is extremely popular in finance. Normal distribution is symmetrical distribution because area in the left and right from the center is same and probability depicted by each area is equal to 0.50 (total probability of 1). The normal distribution is often referred to as the bell curve because of the shape of its probability density function. The normal distribution is the function of mean and standard deviation of the observed data. Please note, normal distribution is bell curve but not every bell curve is normal distribution.

The probability is represented by area under the curve called as probability density function PDF. We use symbol f(x) to represent the curve. Area under the curve is given by a cumulative distribution function (CDF). We don't work with normal distribution because each variable



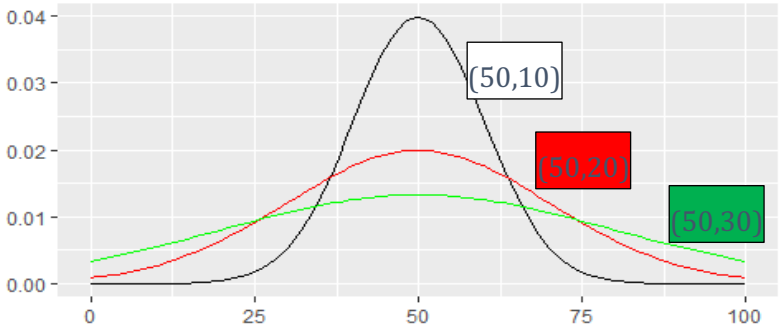
produces its own normal distribution with mean and standard deviation which increase working complexity. The solution for this problem is the converting of normal distribution into Standard Normal Distribution which is very easy to work with because it comes with distribution table called z table providing probabilities which can be applied universally for any normal distribution. We will discuss some properties of normal distribution below which can be applied to standard normal distribution.

The normal distribution has two parameters mean μ and standard deviation σ. Notation for specifying x is normally distributed is $X \sim N(\mu, \sigma)$ and read as x is normally distributed with μ and σ. The probability density function given below for normal distribution is complicated and formula is not important for exam. Cumulative distribution function is P(X<x) i.e. probability of X is less than given value which we don't need to calculate but must be aware of.

The normal distribution curve is perfectly symmetrical with mean = median = mode. The normal distribution is dependent upon the mean and standard deviation which creates shape of the distribution. Smaller the standard deviation narrower the distribution and vice versa.

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Suppose you are provided with a normal distribution with mean of 50 and standard deviation of 10. So obvious meaning of mean here is average of the random variable for the given distribution is 50. Which means 50% of the values of are below 50 and 50% are above 50.



Distribution is measured in standard deviation.

- 1 Standard deviation provides range of 50± 10. Which means values from 40 to 60 are captured by 1 SD.
- 2 standard deviation provides the range of 50±(10 X 2). Meaning 2 unit of standard deviation captures values from 30 to 70.
- 3 standard deviation provides the range of 50± 3X10. Meaning 3 unit of standard deviation captures values from 20 to 80.

This measure offers us a tool to calculate probability of a range for a given distribution. Normal distribution is very well structured with **skewness of zero and kurtosis of 3** (i.e. excess kurtosis of 0). Because of this standardization in shape, the probability captured by 1,2 and 3 SD is fixed.

Table providing probability for 1, 2 and 3 standard deviations – **Empirical rule**

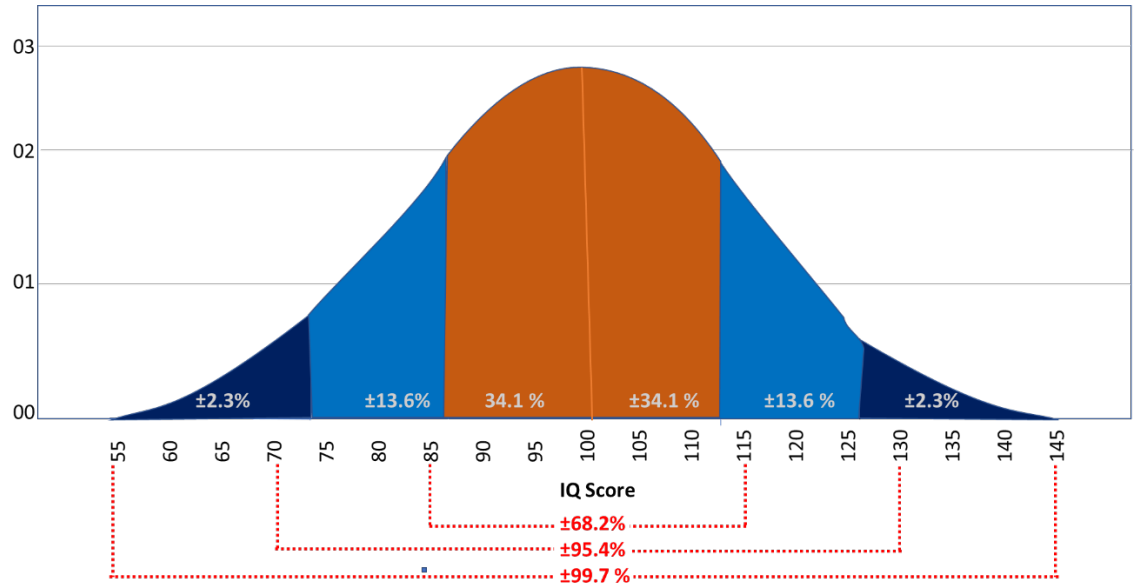
STANDARD DEVIATION	PROBABILITY CAPTURED (TAKING BOTH SIDES)	INTERPRETATION
1 SD	68.2% total probability, which means 34.1% right and left from the mean.	68.2% values of random variable falls in this range. 34.1% values are within 1SD to the left or right of the mean.
2SD	95.4% total, which means 47.7% right and left from the mean.	95.4% values of random variable falls in this range. 47.7% values are within 2SD to the left or right of the mean.
3SD	99.7% total and 49.8% right and left from the mean.	99.7% values of random variable falls in this range. 49.8% values are within 3SD to the left or right of the mean. Almost all the values are captured in ± 3 SD from the mean.

Note: This table is very crucial from the exam perspective, and you must remember all the above values (expect direct questions from this area).

Table providing confidence intervals for commonly used confidence level. Confidence level is the percentage of probability with assures the percentage of values falling in the given range. In the following table provides, for the confidence level of 90% range is mean ± 1.65 SD. This is interpreted as, 90% of the values for a given random variable will fall in between ± 1.65 Standard deviations from the mean.

CONFIDENCE LEVEL	TWO TAIL
90%	Mean ± 1.65 SD
95%	Mean ± 1.96 SD
99%	Mean ± 2.58 SD

Graph showing confidence intervals and probability.



Confidence interval: Confidence interval is the range of values within which a random variable fall with a specific probability. Table given above provides some of the most standard confidence intervals (1 2 and 3 SD). For example, confidence interval of 2 SD is interpreted as 95.45% of the times random variable falls within two SD. These are standard values, but we can find probability for any range like 2.2 SD to 2.5 SD. To deal with non-standard values we need standard distribution (z table) which is covered in the next section. First, we will see how to find probabilities and confidence intervals for the standard values.

CI for given % = $\mu \pm z_{\%} \times \sigma$

Illustration:

Assume students exam score is normally distributed with mean 65 and SD of 15. Using this information and standard set of probabilities provided above, answer the following questions.

QUESTION	CALCULATION
Q1: WHAT IS THE PROBABILITY OF A RANDOMLY SELECTED STUDENT SCORED IN THE RANGE OF 50 TO 65. P(50≤ X ≤ 65)	Step 1: Find the SD range using X-μ /SD. Step2: Find probability. For range 50 to 65 Range = 50 – 65/15 to 65-65/15 = - 1 SD to 0 SD (i.e. mean) This is 1 SD below mean. From the table we know 34.1% probability is 1SD below mean. Hence answer is 34.1%

Q2: WHAT IS THE PROBABILITY OF A RANDOMLY SELECTED STUDENT SCORED IN THE RANGE OF 20 TO 80. P(20≤ X ≤ 80)	20 to 80 = 20-65/15 to 80-65/15 = -3 SD to 1 SD Now we need probability 3 SD below and 1 Sd above mean. 49.8% + 34.1 = 83.9%
Q3: WHAT IS THE PROBABILITY OF A RANDOMLY SELECTED STUDENT SCORED IN THE RANGE OF 20 TO 35. P(20≤ X ≤ 35)	20 to 35 = 20-65/15 to 35-65/15 = -3 SD to -2 SD Now we need probability from -3 SD to -2 SD below mean. 49.8% - 47.7% = 2.1%
Q4: WHAT IS THE PROBABILITY OF A RANDOMLY SELECTED STUDENT SCORED BELOW 35. P(X≤35)	<35 = < 35-65/15 = Less than -2SD. We know probability in the left side from the mean is 50%. To find out probability below 35 simply reduce probability upto -2Sd from the mean from 50%. 50% - 47.7% = 2.3%
Q5: WHAT IS THE 90% CONFIDENCE INTERVAL FOR STUDENTS SCORE.	We know 90% CI gives range of ± 1.65 SD from the mean. Mean ± 1.65 SD = 65 ± 1.65 X 15 = 40.25 to 89.75
Q5: WHAT IS THE 95% CONFIDENCE INTERVAL FOR STUDENTS SCORE.	We know 95% CI is ± 1.96 SD from the mean. Mean ± 1.96 SD = 65 ± 1.96 X 15 = 35.6 to 94.4

Properties of normal distribution

1. Area under the curve is equal to one.
2. Probability is found for intervals of x values rather than for individual x values.
3. Probability of x in continuous random variable is equal to zero (Always).
4. P(a<x<b) is the probability that the random variable x is in the interval between the value a and b.

Key points to remember about normal distribution –

- Skewness of normal distribution is zero
- Kurtosis is 3 and excess kurtosis is 0 for normal distribution. This is known as mesokurtic.

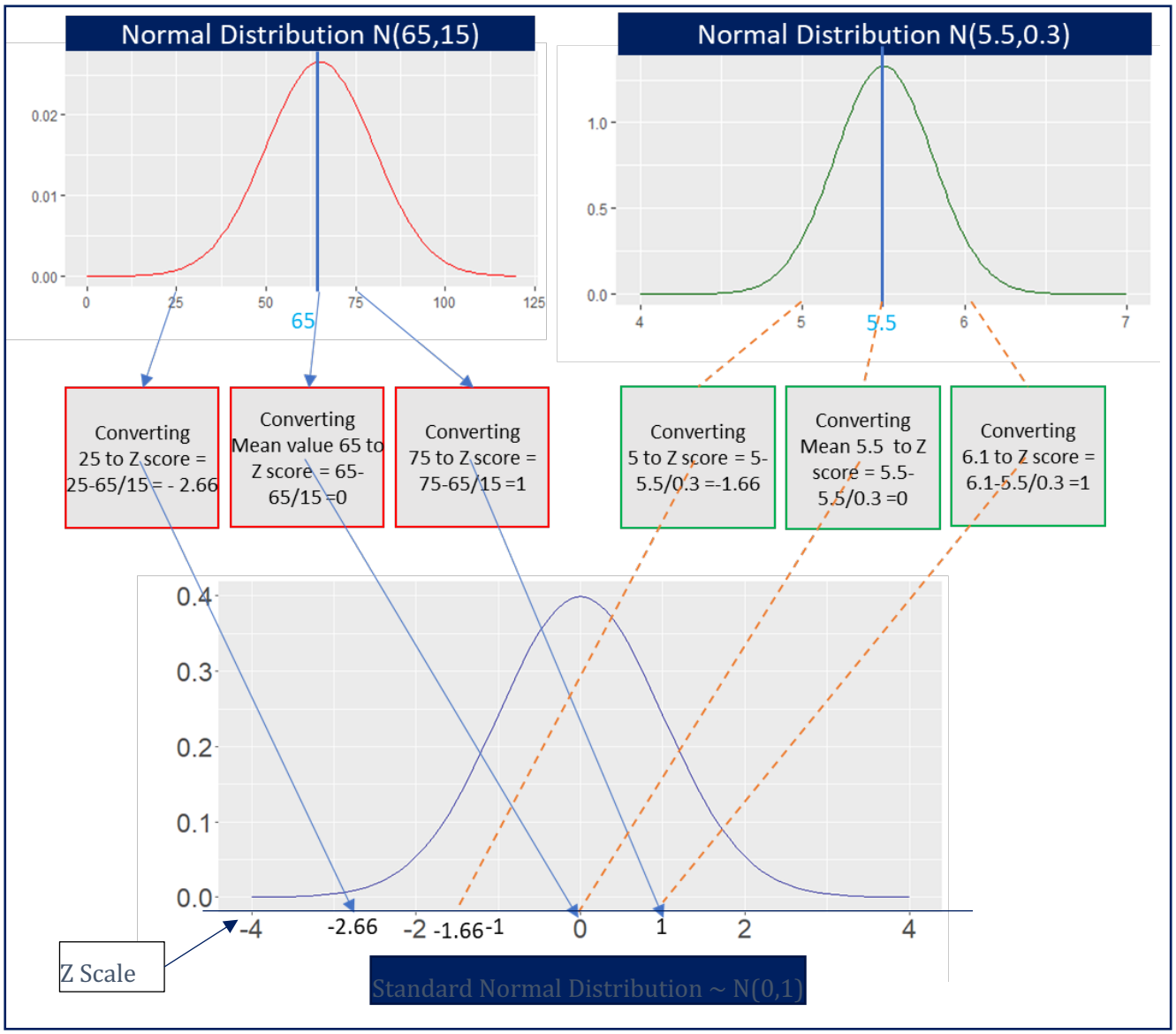
Question: One of the properties mentioned above stated that the probability of x is always equal to zero for continuous random variable. But why?

This is because in continuous random variable for any x value is very miniscule and approximately equal to zero. It is impossible to calculate the probability of exact value of X continuous random variable. Hence, we calculate range as we mentioned in point no 4 above.

Standard Normal Distribution

For every continuous random variable we have different normal distribution because of different mean and SD. Assume you are provided with two normally distributed random variables, Students height ~ N (5.5,0.3), Students exam score ~ N(65,15). In real life we work with many distributions with different means and SD. Things can get overly complicated if we try to create

probability distribution for every continuous random variable. This problem is solved by standard normal distribution. Think of standard normal distribution as a scale to measure probability, which can be applied to any random variable which is normally distributed. Hence, we don't need any separate probability distributions for every continuous random variable. Simply fit the normal distribution to standard normal distribution, and problem solved. Scale of standard normal distribution is called the Z scale and values on this scale are z values or z score. First we will see how can we map a continuous random variable to standard normal distribution. Following is the example of two normal distributions which are mapped to Standard normal distribution.



After this mapping of normal distribution to standard normal distribution, simply use Z table and calculate probabilities.

Standard Normal distribution is also known as the Z-distribution. The total area under the curve is 100%. Like normal distribution we can find the probability with standard normal distribution using area under the curve. The z table is used to find the probability of standard normal distribution. The notation P(z < k) represents the probability of a z-score less than a particular k

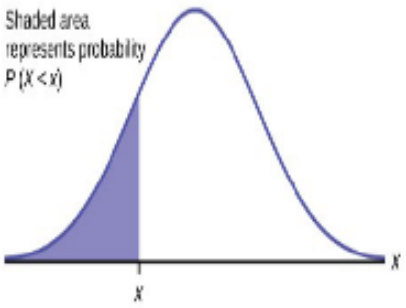
value in the standard normal distribution. Graphical representation of P(z < k). is given below. Example P (z < 0) = 50%.

Formula to standardize random variable into z value /score

$$Z = \frac{Observation\ \ value - population\ mean}{SD} = \frac{x - \mu}{\sigma}$$

For example, z score of 25 for ~ N(65,15) = 25 – 65 / 15 = -2.66 (remember sign is very important here).

Now to find out probability of say P(z < -2.66) or P(x < 25) we need z table.



Z table can be provided in various forms like full table, or partial table with only positive values or negative values and so on. Let's look at two tail table. We can calculate probabilities using any table with slight modification in calculation using same principles. First will start with extract of full table with negative values (probability from left to right).

ENTRY SHOWS P(Z< SPECIFIED Z) -- FOR EXAMPLE:

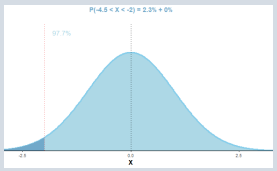
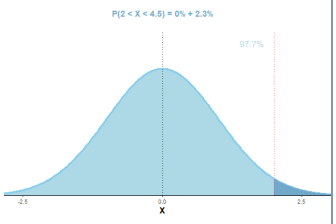
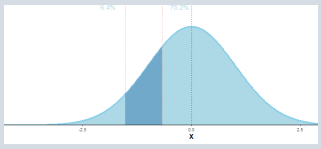
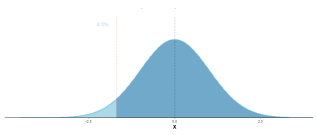
P(Z < 1.24) = .89251

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-2.90	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
-2.80	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
-2.70	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
-2.60	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
-2.50	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
-2.40	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
-2.30	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
-2.20	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
-2.10	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
-2.00	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831

Z table is provided in the matrix format. First column of this table provides z value upto only 1 decimal. First row provides 2nd decimal value of column. To find Z value probability of say 2.42 – find 2.4 in column and 0.02 in row, and then find intersection point (color coded in table). This table provides probability of area below z value p(z< z specified).

Getting back to our above illustration to find P(z < -2.66) will find the intersection value of 2.6 and 0.06 which is 0.00391. Hence probability of x less than 25 is 0.00391 or 0.391%. This concept is further explained with following illustrations

Illustration: Find out various probabilities for given ranges.

Q1: P(z<-2.00)		From the table 0.02275
Q2: P(z> 2.00)		Recall, Standard ND is symmetrical, hence area above 2.00 is equal to below – 2.00. From the table 0.02275
Q3: **P(-2.8 <z < -2.1)**		Area < -2.8 = 0.00256 Area < -2.1 = 0.01786 P(-2.8 < z < -2.1) = 0.01786 – 0.00256 = 0.0153
P(z> -2.55)		First find the probability of area below – 2.55 which is= 0.00539. We know total area is 1 and below -2.55 is 0.00539, hence above -2.55 is 1 – 0.00539 P(Z>-2.55) = 1-0.00539 =0.9946

We can also find z value for given probability using table. Example, what is the z value to cover 2% of lowest values? To answer this question first find 2% i.e. 0.02 probability in z table and respective z value is the answer. We don't have exact 0.02 in z table. Closest values of probability are 0.02018 and 0.01970 for z value -2.05 and -2.06 respectively. We can find z value for probability 0.02 by using linear interpolation or simply taking average. Even if we take simple average of -2.05 + (-2.06) / 2= 2.055 is good approximation in this case.

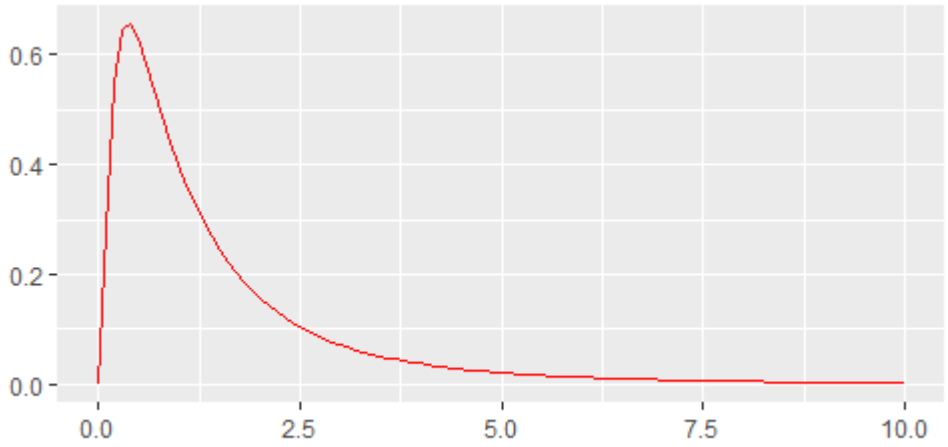
3.3.b The lognormal distribution

We learned in the previous section that; normal distribution is symmetrical with positive or negative values extending to infinity. Distribution functions are close ended mathematical functions, which means irrespective of the actual distribution of the random variable, which required inputs it produces result. Assume random variable is not normally distributed and we use mean and standard deviation of random variable as an input in normal distribution function, it will produce resulting probabilities (which is wrong results). If we have random variable which do not assume any negative values, using normal distribution function is not the correct choice. We want some distribution which can restrict values to positive only (i.e. bound by zero). Take the example of share prices, lowest share price can be 0 and can never take negative value. The solution to this problem is lognormal distribution which is **bounded by zero** (never take negative

value) and **positively skewed** distribution. Stock price may be well described by the lognormal distribution when stock returns are normally distributed (and even if stock returns are not normally distributed). Logarithms of lognormally distributed random variables are normally distributed ($\ln(e^x) = x$.)

We can also apply this concept to stock returns. We very well know that lowest (worst) possible stock return is -100%. Assume you own a stock worth $500 and company goes bankrupt next day, which leads to stock worth $0. This is 100% loss hence worst loss is -100%. But normal distribution values extend to infinity in both ends. Hence modeling stock prices using lognormal distribution is better approach.

Statement to remember: If log of returns are normally distributed then one plus standard returns (1+r) are lognormally distributed.



3.3.c Student's t Distribution

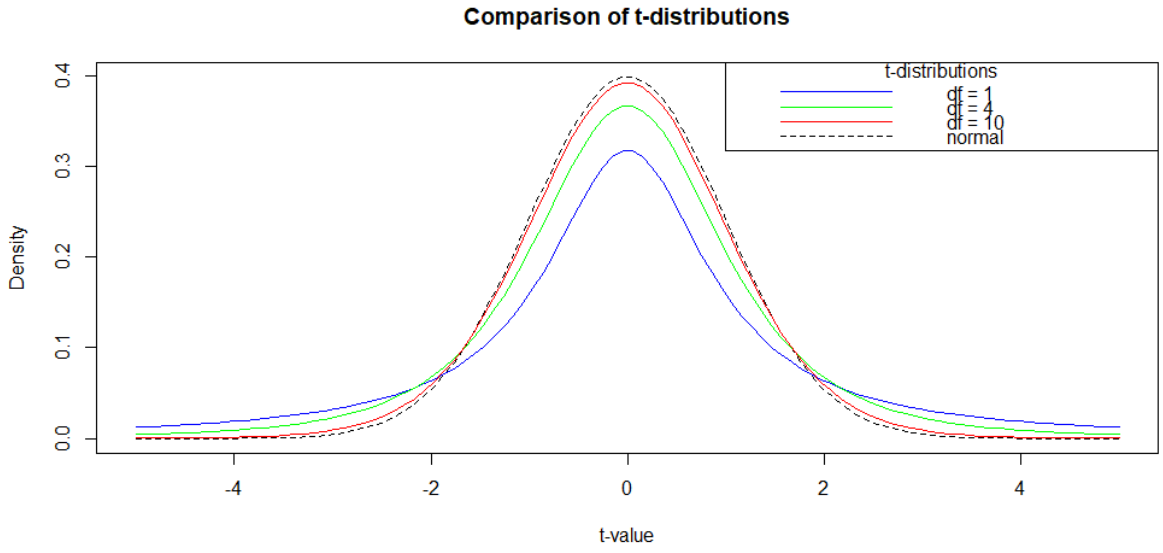
Another extremely popular distribution in statistics and in risk management is Students t - distribution. This distribution has nothing to do with Students, it is just pseudonym used by William Gosset to publish the paper on distribution which works better with small sample sizes. To build the confidence interval using normal distribution, population standard deviation should be known. Population parameters are rarely known in practice. When the sample size is large enough this does not create any problem in estimating population parameters using sample estimates. However, Gosset observed, when sample size is small with unknown variance, it results into inaccuracy in confidence interval. To tackle this problem Student's t distribution was created which do not take support of variance to construct confidence intervals.

Properties of students t distribution

- Student's t-distribution is symmetrical with mean 0, like the standard normal distribution; however, it has **more probability in its tails than the standard normal distribution**.
- t-score is used in t-distribution which is analogues z score of Standard normal distribution (with same meaning and interpretation).
- It is defined by single parameter, the degree of freedom (df), where df = n – 1. (n = number of observations/ samples).

Table showing t score for given df and probability in upper one tail t distribution.

	ONE TAIL P =0.1	P =**0.05**	P =**0.025**	P =**0.01**	P = **0.005**
DF	Two tail p = 0.20	P = 0.10	P= 0.05	P=0.02	P=0.01
1	3.07768	6.31375	12.70620	31.82052	63.65674
2	1.88562	2.91999	4.30265	6.96456	9.92484
3	1.63774	2.35336	3.18245	4.54070	5.84091
4	1.53321	2.13185	2.77645	3.74695	4.60409
5	1.47588	2.01505	2.57058	3.36493	4.03214
6	1.43976	1.94318	2.44691	3.14267	3.70743
7	1.41492	1.89458	2.36462	2.99795	3.49948
8	1.39682	1.85955	2.30600	2.89646	3.35539
9	1.38303	1.83311	2.26216	2.82144	3.24984



Observations from the above diagram

- As the degrees of freedom increases t-distribution's peak increases
- Lower the degrees of freedom means more probability in the tails.
- T distribution converges to normal distribution (dotted line) as df increases.

Building confidence interval using students t distribution is same as normal distribution, just replace z value and standard deviation in mean ± z x SD with t value and standard error mean ± t x SE.

Where SE = $Standard\ error\ = \frac{s}{\sqrt{n}}$

Note: We will see use case of this distribution in Reading 06 Hypothesis Testing. Following distributions – Chi squared and F-distributions can be understood in better manner with the help of Reading 6 Hypothesis testing, which was removed from FRM curriculum since 2020 curriculum updates. Hence, we will only take the overview of these two concepts.

3.3.d Chi squared distribution

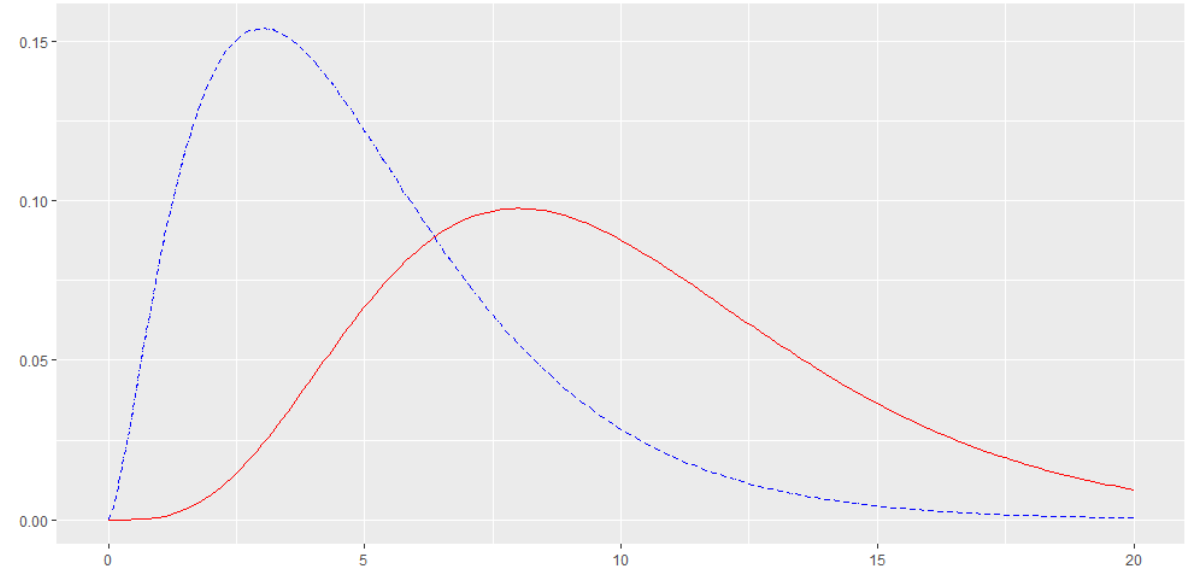
If we have k independent normal variables, Z1, Z2, …, Zk, then sum of their squares, S, has chi squared distribution. K is degrees of freedom. Because Chi squared variable is the sum of squared values (hence the name Chi (K) squared), it only assumes non-negative values and is asymmetrical. The mean of the distribution is k and variance is 2k. As the k increases, distribution becomes symmetrical. We are not concerned with the density function of the chi squared distribution. We need it for hypothesis testing and hence the only concern from the exam perspective is chi squared statistics.

Chi squared test statistics with n – 1 degrees of freedom, is computed as

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

Where n is sample size, S^2 is sample variance, and σ^2 is hypothesized value of population variance.

Blue dotted line df = 5, Red line df = 10



3.3.e F Distribution

F distribution is used in testing of two variances simultaneously. It is often desirable to compare two variances rather than two averages. For example, college administrators would like two college professors grading exams to have the same variation in their grading.

To perform an F test of two variance, it is important that the following are true:

- The population from which the two samples are drawn are normally distributed.
- The two population are independent of each other.

F test for equality of two variance is extremely sensitive to deviation from normality. If the two distributions are not normal, the test can give higher p-value that it should, or lower ones, in ways that are unpredictable. Suppose we sample two independent normal populations. Let σ_1^2 and

σ_2^2 be the population variance and s_1^2 and s_2^2 be the sample variances. Let the sample sizes be n1 and n2. Since we are interested in comparing the two sample variances, we use the F ratio.

$$F = \frac{s_1^2}{s_2^2}$$

where,

S_1^2 = variance of the sample of n1 observation drawn from population 1.

s_2^2 = variance of the sample of n2 observation drawn from population 2.

Properties of F distribution

- All F values are greater than or equal to 0
- There is a different F curve for each pair of degrees of freedom $n_1 - 1$, n_2-1.
- Curve is nonsymmetrical and skewed to the right.
- There is 100% under the curve.

Relation between the F and Chi squared distribution such that:

$$\boldsymbol{F = \frac{\chi^2}{\#of\ observation\ in\ numerator}}$$

3.3.f The Exponential Distribution

The exponential distribution is often concerned with the amount of time until specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution. Other examples include time taken by a bank to default and the amount of time in months a phone battery lasts.

Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money.

Exponential distribution is closely related to the Poisson distribution. Poisson distribution gives the probability of x as a result in specific time interval. Exponential distribution gives the time interval for x as a result. Example, Poisson distribution – Probability of total 2 companies default in two years. Whereas, example of exponential distribution - Time taken by a company to default.

PDF of exponential β is $f(y) = \frac{1}{\beta} X\ e^{-\frac{y}{\beta}}$ where, β is $\frac{1}{\lambda}$,

Exponential variables are also memoryless, meaning their distribution are independent of their histories. Example, company default probability for the first year is same as the default probability of second year. If company doesn't default in first year it does not increase the probability of default in second year. This is called as memoryless. This does not imply that the probability of company default in first year is equal to company default in first two years.

Illustration No: 3.4

Assume that the time to default for a consumer loan is exponentially distributed with β of 2 years. Find the probability that consumer will default within 3 years.

In the above illustration β = 2 and y = 3 years.

Default in 3 years = 1- $e^{-3/2}$ = 0.7768 = 77.68%.

3.3.g Beta distribution

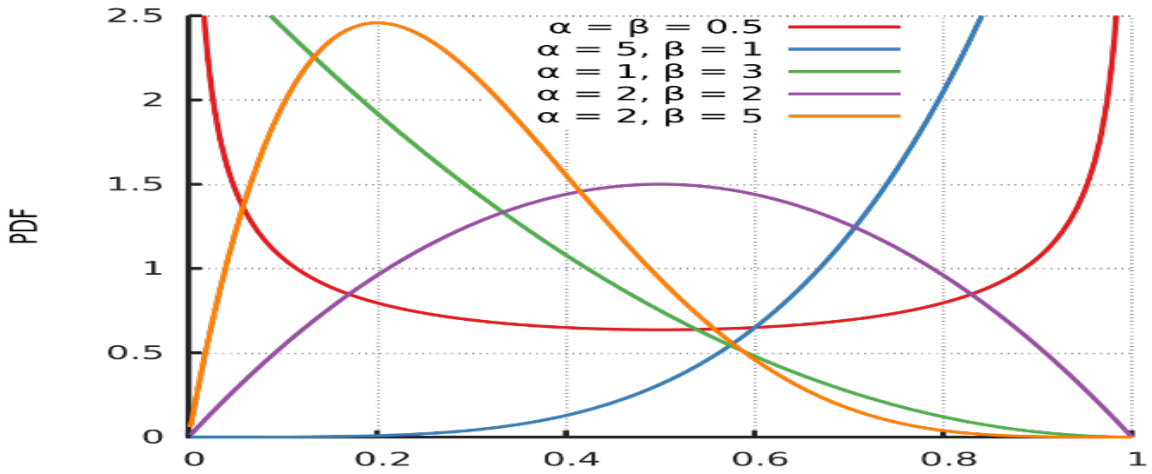
The beta distribution applies to continuous random variable with outcome between 0 and 1. It is commonly used to model probabilities that naturally fall into this range. The beta distribution has two parameters, α and β., that jointly determine the mean and variance of a random variable which is Beta distributed. If Y~ Beta(α,β)

$E[Y] = \frac{\alpha}{\alpha+\beta}$

And $V[Y] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Properties of Beta Distribution

- Distribution places most of the probability mass near the boundaries when both α and β < 1.
- Distribution is standard uniform distribution when α = β = 1.
- As the parameters increases above 1, distribution becomes more concentrated around the mean.



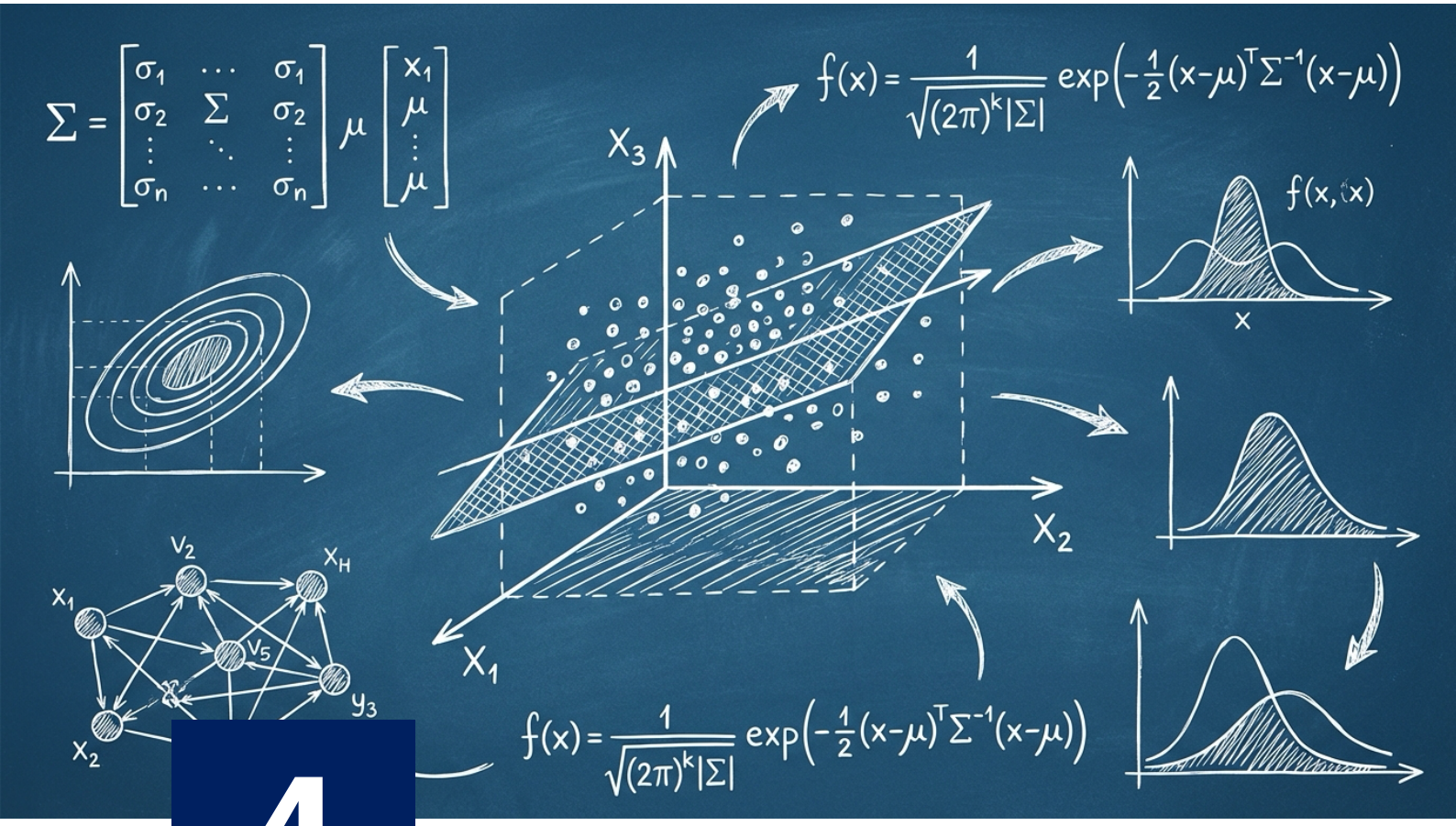
3.3.h Mixture Distribution

Mixture distributions build new, complex distribution using two or more component distributions. A two-component mixture first draws a value from a say Bernoulli random variable distribution (simple distribution of 0 and 1). Then depending upon value 0 or 1, draws from one of two component distributions.

Properties of mixture distribution

- Both PDF and CDF of a mixture distribution are the weighted averages of the CDFs and PDFs of the component.
- Mixture distribution can have both skewness and excess kurtosis even when their components have no skewness or excess kurtosis.

Reading 4 Multivariate Random Variables



4 Multivariate Random Variables

SCOPE OF THIS READING

This chapter develops the framework for analyzing discrete bivariate random variables. It explains how a probability matrix represents a joint probability mass function and how to derive marginal and conditional distributions. The chapter formalizes expectation for functions of two random variables and introduces covariance and correlation as measures of linear dependence, linking them to independence. It analyzes the impact of linear transformations on covariance and correlation, derives the variance of a weighted sum, and computes conditional expectations. Finally, it defines the properties of independent and identically distributed (IID) sequences and explains how the IID assumption simplifies the computation of the mean and variance of sums of random variables.

Note: Multiple learning objectives from this reading are covered in Level 0 Reading Basic Statistics.

4.1 Applying Linear Transformation on Covariance and Correlation between two random variables

Correlation measures the strength of the linear relationship between two variables and is always between -1 and +1. Linear transformation on correlation and covariance between two variables works in defined manner.

Note: Concept discussed below takes support of linear regression which is covered in Reading No 07. For now, just focus on basics of these concepts, once you study linear regression topic you will understand reasoning behind it.

If $X_2 = a + b X_1$, then correlation between X_2 and X_1 is

- 1 if b>0
- -1 if b < 0
- 0 if b = 0

This can be directly verified using correlation formula, but we don't need it for exam purpose, hence we can skip verification part.

In the Reading 2 and level 0 basic statistics we learned variance of $a + bX_1$ is b^2 var(X_1). This means that 'a' in this equation shifts location(mean) by 'a' and have no effect on variance, while rescaling by b scales the variance by b^2.

Applying same principle on covariance of two random variables X_1 and X_2.

Cov $(a+bX_1, c + dX_2) = bd$ Cov(X_1,X_2)

In the above case location is unaffected and scale of each component is affected by b and d multiplicatively. Combining these two properties, we can infer that the correlation is unaffected by scale (scale free).

Corr$(a+bX_1, c+dX_2) = \frac{bd\ Cov(X1,X2)}{bd\ Sigma\ X1\ Sigma\ x\ 2}$ = sign(b) sign(d) Corr(X1,x2)

Coskewness and cokurtosis: Like skewness and kurtosis in one variable, coskewness and cokurtosis are cross variable versions for two random variables are also standardized. Interpretation of coskewness and cokurtosis is not very clear.

4.2 the variance of sum of random variables

The covariance is important in calculation of variance of two random variables.

$V(X_1 + X_2) = V(X_1) + V(X_2) + 2$ Cov (X_1,X_2)

And $V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) + 2ab$ Cov(X_1, X_2) – equation (a)

This property plays very important role in portfolio construction.

Note: Topic discussed below is directly related to Book 1 CAPM reading. At this point if you are unable to grasp this concept don't worry, when you cover CAPM reading you will get the meaning and purpose of this concept.

a and b given in above equation (a) acts like weight of the asset. In the two-asset portfolio application, we can calculate the variance of portfolio.

Example

From our equation	*Example*	*Values*
a	Weight of asset	X1 = 0.60
b	Weight of asset	X2 = 0.40
X1	Return of asset X1	SD of X1 = 0.12
X2	Return of asset X2	SD of X2 = 0.08

Note: X_1 and X_2 are return of assets. We are taking SD of X_1 and X_2 in value column instead of return (which will be series of returns, using which we can calculate SD) to ease our calculation.

SD (w_1X_1,w_2X_2) = $\sqrt{\left(w_1 SD(X1)\right)^2 + \left(w2SD(X2)\right)^2 + 2\, w1w2Cov(X1X2)}$

The optimal weight for lowest variance can be shown to be

$W^* = \frac{\sigma 22 - \sigma 12}{\sigma 11 - 2\sigma 12 + \sigma 22}$

For calculation purpose we will assume the covariance of 0.005568

Calculation of SD(X_1,X_2) = $\sqrt{(0.60 * 0.12)^2 + (0.40 * 0.08)^2 + 2 * 0.60 * 0.40 * 0.005568}$ = 0.1194

Hence the standard deviation of two asset portfolio with given weight is 0.1194

4.3 Independent and identically distributed random variable

This assumption requires, two properties to be fulfilled by two random variables to be considered as IID (independent and identically distributed)

- Both the random variables are independent. i.e. probabilities are independent (like in coin toss experiment)
- Both the random variables have identical distribution, or we can also say both are drawn from the same distribution (Identical in mean and SD i.e all RV with same mean and sd).

This is very important property. If the events are independent, then correlation between two random variables is zero. This gives us following solutions.

E ($\sum$Xi) = n μ_i

Var ($\sum$Xi) = n SD^2

In the absence of zero correlation, for the variance calculation (when correlation and covariance exists) we will need n(n-1) / 2 distinct pairs. Zero correlation results in simplification of this equation.

Please note the important distinction between the variance of sum of multiple random variables and variance of multiple of a single random variable.

$V(X_1+X_2) = 2SD^2$

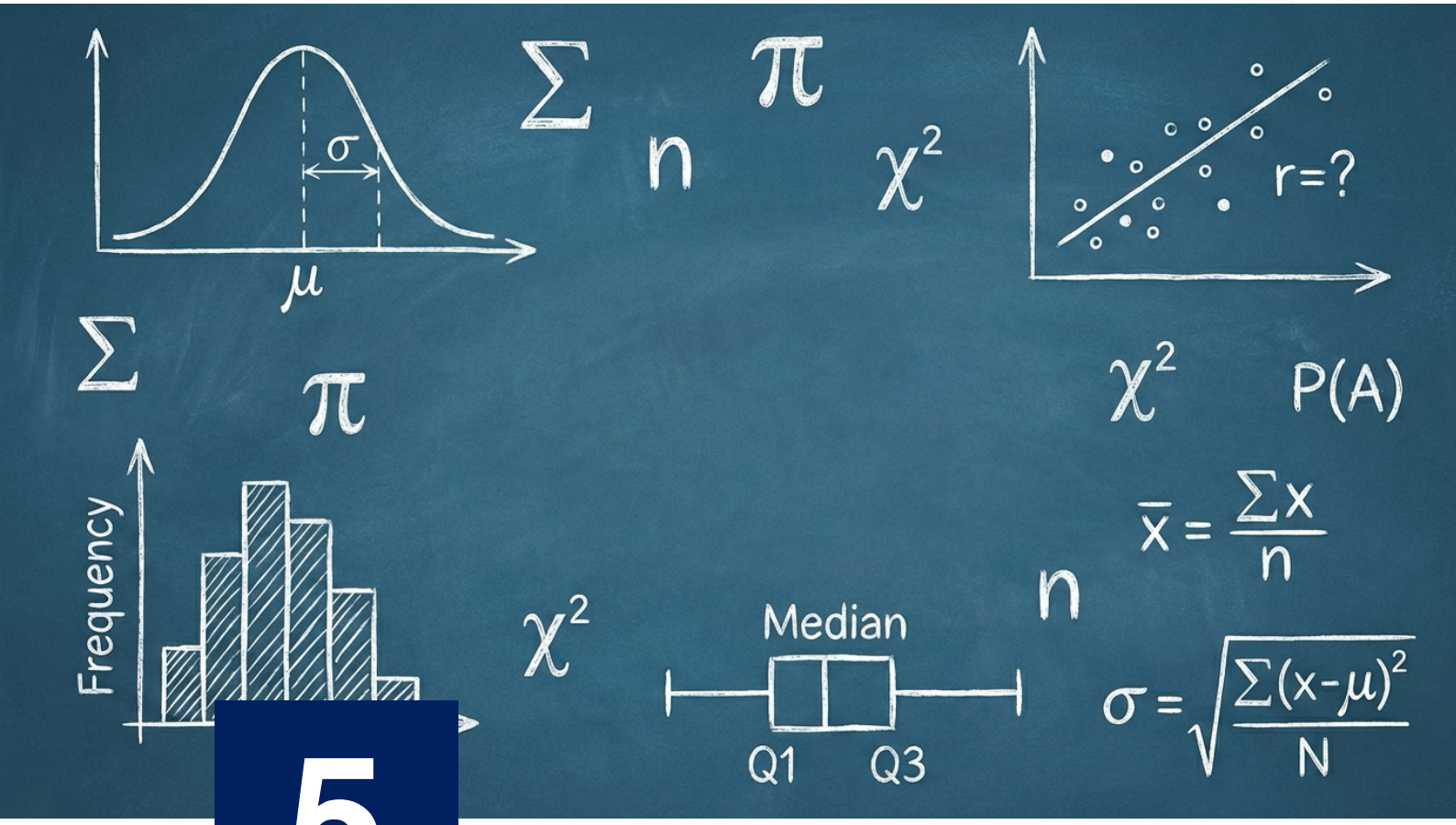
Is different from

$V(2X_1) = 4\ SD^2$

$V(2X_2) = 4\ SD^2$

This property plays an important role when estimating unknown parameters. The variance of the sum of iid random variables grows linearly. This means that when the sum of n random variables is divided by n to form an average, the variance of the average reduces as n grows.

Reading 5 Sample Moments



5 Sample Moments

SCOPE OF THIS READING

This chapter develops statistical estimation using sample data. It explains how to estimate mean, variance, standard deviation, skewness, kurtosis, quantiles, covariance, and correlation, and distinguishes between population moments and sample moments, as well as between an estimator and an estimate. The chapter analyzes properties of estimators, including bias, consistency, and the conditions under which the sample mean is BLUE. It applies the Law of Large Numbers and the Central Limit Theorem to sampling distributions, extends estimation to joint means of two variables, and introduces higher-moment comovements through coskewness and cokurtosis.

Note: Multiple learning objectives from this reading are covered in Level 0 Reading Basic Statistics.

5.1 Point Estimate and Estimator

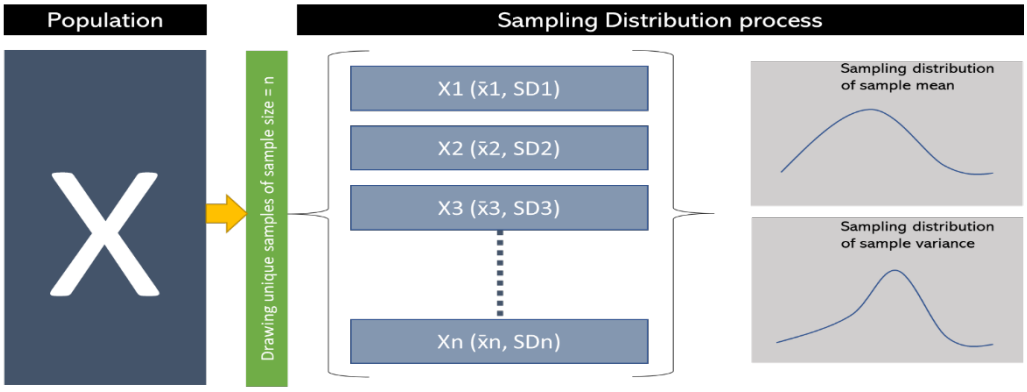
Collecting data from a whole population for analysis is very costly and often not feasible. Therefore, we use sample data taken from the population to reduce the cost and make it possible to do. The aim here is to analyze and infer based on sample which reflects the population data. Using sample data can have a lot of risks. Choosing unsuitable data sampling methods known as biased sampling methods can lead to erroneous conclusions about the population. Sampling methodology is not covered in FRM curriculum in depth. Another problem in using sample data is selecting correct formulas for estimating moments like mean and SD based on samples data which represent population parameters. Sometimes we need to adjust population parameters formula to get the better estimation of moments using sample data. Population can be best explained by population parameters (i.e., mean, SD etc calculated using population data). Because populations are often very large or hard/expensive to examine, mostly we have no way to know the exact values of parameters (like true mean and SD).

The point estimators are used to infer population parameters. The value of the estimator is called an estimate. For example, if the sample data has a mean of 50, then the mean is an estimator of the population parameter and 50 is the estimate, which is the value of the mean. Estimates vary depending on how the samples are drawn. Estimators vary depending on the function used and so there can be different estimators for the same population parameters. Therefore, we need to find good estimators (which bring the sample estimates close to the population data). To assess the quality of an estimator, it is important to know things about the estimator's sampling distribution, its mean, its variance etc.

5.2 What is Sampling distribution?

Sampling distribution is the term for the distribution of sample estimates based on the population. Sampling distribution is the distribution of all the different sample estimates that are randomly taken from the same population. We can create sampling distribution of mean and sampling distribution of standard deviations (see fig below). Sampling distribution of mean is the distribution of the means that are computed using samples taken, and sampling distribution of sample variance is the distribution of the variances that are computed using random samples.

Term	**Description**
Sampling distribution	Distribution of sample estimates based on the population
Sampling distribution of mean	Distribution of the means computed using samples taken
Sampling distribution of sample variance	Distribution of the variances computed using random samples



Properties of sampling distribution

- The mean of sampling distribution of x-bar is equal to the population mean.
- The standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$, where n is the sample size. Hence sample means distribution is N(μ, $\frac{\sigma}{\sqrt{n}}$,).
- Standard deviation of sampling distribution is the standard error of mean of sampling distribution.
- For normally distributed population, sampling distribution is also normally distributed.
- For non-normal population, sampling distribution is normal if sample size is large enough (usually ≥30). This statement takes the support of Central limit theorem (discussed later in this chapter).

More on sampling error: When we draw samples from the population, mean of sample and mean of population is rarely same. Hence drawing inference about the population based directly on sample is not at all a good idea. This problem is solved by sampling distribution. Mean of sampling distribution on the other hand is close to population mean. But there is still some inaccuracy left. This inaccuracy is measured by sampling error i.e. standard deviation of the sampling distribution. Hence the ***goal is to have sampling error as low as possible***, so that population can be estimated more accurately. Sampling error is very useful in hypothesis testing which we will study in the next chapter. Sampling error is the function of standard deviation and sample size n. When it comes to keeping standard error low, we don't have control over standard deviation of sampling distribution, but we can increase sample size n which will lower sampling error.

5.3 Bias of an estimator and bias measures

For statistical analysis we are interested in the value of population parameters such as the mean or the variance. However, these values are not observable for obvious reasons, and so sample data is used to estimate these values. Estimators may have some difference between the expected value of the estimator $E[\hat{\theta}]$ and the true population value θ. This difference is called estimator bias. Following table provides the summary of two main estimators and their biases. Reasoning behind it is not very important for exam and bit complicated hence not discussed here (please ref GARP book for reasoning in case you are interested in it).

EXPECTED VALUE OF	BIAS (WHEN IID)	IS BIASED?	BIAS CALCULATION
MEAN	$Bias(\hat{\mu}) = E[\hat{\mu}] - \mu$	Unbiased Estimator	$\mu - \mu = 0$
SAMPLE VARIANCE	$Bias(\hat{\sigma}^2) = E[\hat{\sigma}^2] - \hat{\sigma}^2$	Biased estimator	$\frac{\sigma^2}{n}$

Important note: We know the bias in sample variance. With the help of this bias, we can determine unbiased estimator of sample variance.

Unbiased estimator of variance: This is same formula we used for variance calculation using sample data in basic statistics chapter. This is also reasoning behind using n-1 in the denominator for calculation of variance using sample data, i.e. to make variance estimator unbiased.

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

5.4 blue – best linear unbiased estimators

The mean estimator is Best Linear Unbiased Estimator (BLUE) of the population mean when the data are iid, because mean estimator has **the lowest variance** of any linear unbiased estimator. BLUE is a desirable property for an estimator, because it establishes that the estimator is the **best estimator among all linear and unbiased estimator**. It does not mean that there are no better estimators to the sample mean, but those are not linear. Like maximum likelihood estimator of the mean is generally more accurate tahn the sample mean, but it is not linear and often biased in finite samples. We prefer linear estimator over non-linear estimator for ease of calculation.

5.5 Law of large numbers (LLN) and Central Limit Theorem

The law of large numbers establishes the large sample behaviour of mean is similar and provides the condition where the mean converges to its expectation. The simplest element for iid random variables is the Kolmogorov Strong Law of large Numbers. *LLN provides a set of sufficient conditions for convergence of the sample mean to the constant which is equal to expected value of the distribution. LLN states some conditions that are sufficient to guarantee this convergence as the sample size n increases.*

Implications of LLN is consistent estimator:

- When LLN applies to an estimator, the estimator is said to be consistent. Consistency requires that an estimator is asymptotically unbiased (bias of the estimator approaches zero as n tends to infinity), and so any finite sample bias must diminish as n increases.
- As the n increases variance of the estimator converges to zero.

Assumptions in LLN:

- Mean is finite.

5.5.a CLT - Central Limit Theorem

In simple terms, CLT states that, for the large samples size of n, the distribution of the sample means drawn from the population with mean μ and variance σ^2 will be approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$.

CLT extends LLN, provides an approximation to the distribution of the sample mean estimator. Furthermore, they do not require knowledge of the distribution of random variables generating the data. In fact, only independence and some moment conditions are required for CLT to apply to a sample mean estimator. CLT is used as an approximation in the finite sample so that the distribution of the sample mean is approximated. For the mean in large sample the distribution of the sample mean estimator is cantered on the population mean and the variance of the sample average declines as n grows.

Assumptions in CLT:

- Mean is finite (same as LLN)
- Variance is finite (additional assumption compared to LLN)

CLT does not require assumption about the distribution of the population because for the large n sampling distribution is normally distributed.

Summary of CLT properties:

- If sample size is large enough (n ≥ 30), sampling distribution is approximately normal.
- The population mean and the sampling distribution mean are equal.
- Variance of the sampling distribution is $\frac{\sigma^2}{n}$ and approaches to zero as sample size increases.

5.6 Mean of the two random variables

We can estimate mean of the two random variables in the same manner as we do for the single random variable.

$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i$

When the data are iid, the CLT applies to each estimator by stacking the two mean estimators into a vector. $\hat{\mu} = \begin{bmatrix}\hat{\mu}_X \\ \hat{\mu}_Y\end{bmatrix}$

This vector is asymptotically normally distributed if multivariate random variable z = [x,y] is iid. This assumes each component has a finite variance.

In the bivariate CLT, correlation in the data produces a correlation between the sample means and correlation between the means is identical to the correlation between the data series.

5.7 Coskewness and Cokurtosis

Skewness and kurtosis can be extended to pairs of random variables. When computing cross p^{th} moments, there are p – 1 different measure.

MOMENT	NUMBER OF CROSS MOMENTS
MEAN	Zero cross moment
VARIANCE	One cross moment. Covariance
SKEWNESS	Two cross moments. Coskewness
KUTROSIS	Three cross moments. Cokurtosis

5.7.a Coskewness measures

Two Coskewness measures are

$$S(X, X, Y) = \frac{[E(X - E[X])^2(Y - E[Y])]}{\sigma_x^2 \sigma_Y}$$

$$S(X, X, Y) = \frac{[E((Y - E[X])(Y - E[Y])^2]}{\sigma_x \sigma_Y^2}$$

Coskewness like skewness is standardized version and hence it is scale and unit free. These measures capture the likelihood of the data taking a large directional value, whenever the other variable is large in magnitude. When there is no sensitivity to the direction of one variable to the magnitude of the other, the Coskewness is zero.

5.7.b Cokurtosis measures

Cokurtosis uses the combination of powers that add to 4 with three possible combinations.

$$k(X, X, Y, Y) = \frac{[E((X - E[X])^2(Y - E[Y])^2]}{\sigma_X^2 \sigma_Y^2}$$

$$k(X, X, X, Y) = \frac{[E(X - E[X])^3(Y - E[Y])]}{\sigma_X^3 \sigma_Y}$$

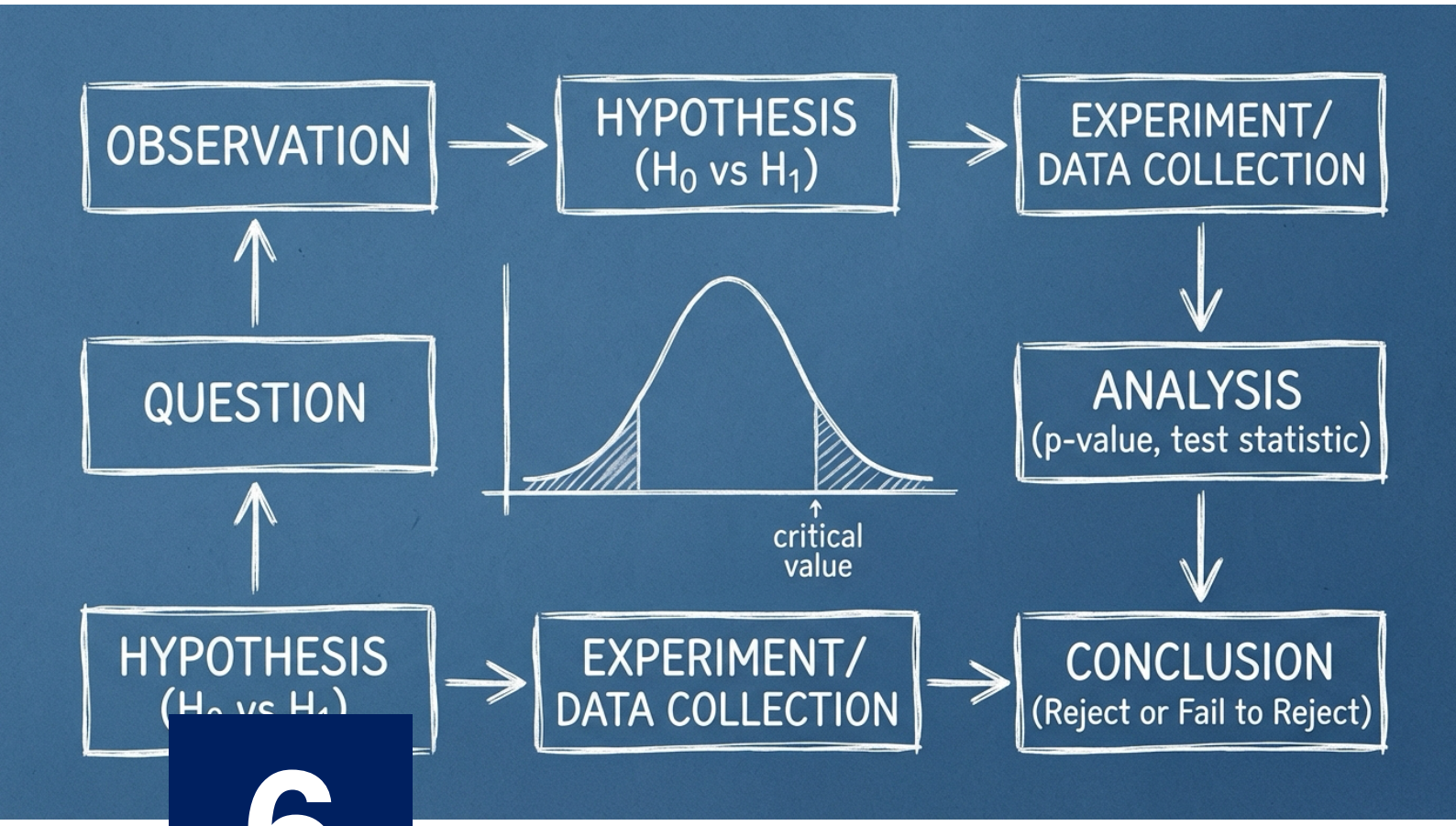
$$k(X, Y, Y, Y) = \frac{[E((X - E[X])(Y - E[Y])^3]}{\sigma_Y \sigma_Y^3}$$

When examining kurtosis, the value is usually compared to the kurtosis of a normal distribution = 3. Comparing Cokurtosis to that of a normal distribution is more difficult, because the Cokurtosis of a bivariate normal depends on correlation.

Points to remember:

- The **symmetric Cokurtosis k(xxyy)** always ranges between **1 and 3.** It is 1 when correlation is zero and rises symmetrically as the correlation moves away from 0.
- The asymmetric kurtosis ranges between -3 to 3 and is linear in correlation.

Reading 6 Hypothesis Testing



6 Hypothesis Testing

SCOPE OF THIS READING

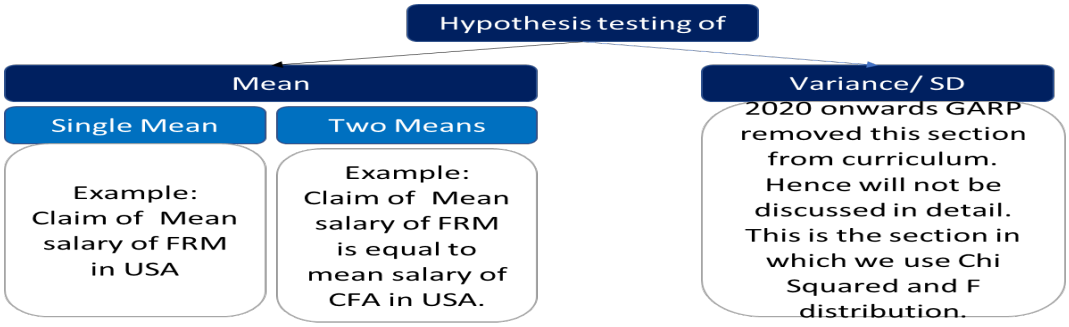
This chapter develops the framework of statistical hypothesis testing. It explains how to formulate null and alternative hypotheses and distinguishes between one-sided and two-sided tests. The chapter analyzes Type I and Type II errors and relates them to test size and power. It clarifies the interpretation of p-values and the relationship between hypothesis tests and confidence intervals, including construction and interpretation at specified confidence levels. It also outlines procedures for testing differences between two population means and examines the problem of multiple testing and its implications for biased inference.

6.1 Introduction

What is hypothesis and hypothesis testing?

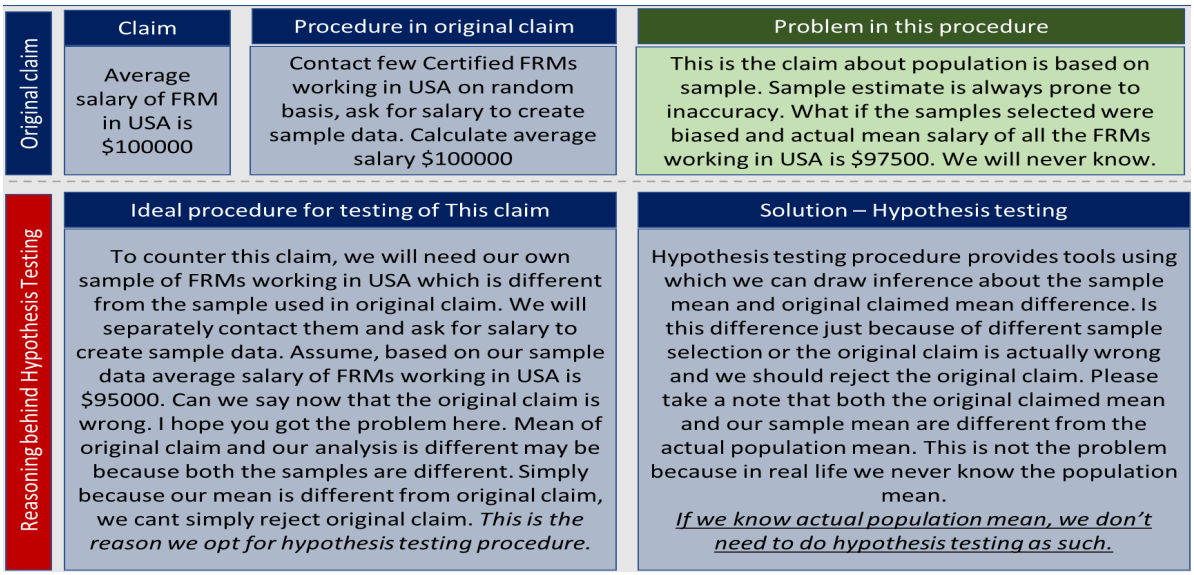
One of the job of statisticians is to make statistical inferences about the population based on sample taken from the population. Assume, a research institute published article on global demand for Certified FRM's. Article stated that the average salary of FRMs working in USA is $100000 PA. This research is produced using the sample data, which makes it subject to error. If it had been created using population data, this would not have occurred. Assume, you disagree with this information on average salary, because in our opinion, salary is understated, and this error is due to sample selection. Being statistician, to reject this claim we need to follow procedure called hypothesis testing. Hypothesis in simple words means a claim of something to be true. In our example, mean salary of FRMs is hypothesis. Procedure opted to check the validity of this claim is called as hypothesis testing.

A hypothesis test involves collecting sample data and evaluating the data to decide as to whether or not there is sufficient evidence based upon sample data analysis, to reject the hypothesis. Hypothesis testing can be conducted on any population parameter but most common in hypothesis testing are mean and standard deviation.



Why do we need hypothesis testing procedure, can't we reject this claim simply based on values estimated from our analysis?

To answer this question first we need to understand the sequence of events starting from the original claim. Look at the following chart to understand the whole process.



Outline of Hypothesis Testing process

Step 1: Set up two contradictory hypotheses: Hypothesis testing starts with setup of Hypothesis. Original claim (mean salary of FRM) is called null hypothesis and statement used to counter this claim is alternate hypothesis.

HYPOTHESIS	WHAT IS IT?	EXAMPLE	STATEMENT
H_0: THE NULL HYPOTHESIS	Hypothesis statement about the original claim	H_0: Average salary of FRM is $100000	$H_0: \mu = 100000$
H_A: THE ALTERNATIVE HYPOTHESIS	Alternative statement to counter this claim	H_a: Average salary of FRM is not equal to $100000	$H_a: \mu \neq 100000$

Step 2: Collect sample data: In this step we will create our own list of contacts of FRMs working in USA. Will ask them salary and create our sample data set.

Step 3: Identify the appropriate test statistics and distribution to perform hypothesis testing: Depending upon the test scenario, we must select right distribution. This depends upon the parameter which we are testing (i.e. mean or SD) and sample size (discussed in detail in later part of this reading).

Step 4: Specify significance level: Hypothesis testing is conducted at specific confidence level. In simple words, say 95% confidence level, we are 95% confident about the decision of hypothesis.

Step 4: Find out the sample statistic and t or z critical value which will be ultimately used to reject or fail to reject the null hypothesis.

Step 5: Decision Making: Reject null or fail to reject null.

6.2 Null and Alternative Hypothesis

The actual test begins by considering two hypotheses, the null hypothesis, and the alternative hypothesis. These hypotheses contain contrary viewpoints.

H0: The null hypothesis: It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.

Ha: The alternative hypothesis: It is a claim about the population that is contradictory to H0 and what we conclude when we reject H0.

Hypothesis test: The aim in a hypothesis test is to decide whether the null hypothesis should be rejected in favor of the alternative hypothesis.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data collected separately for this purpose.

Then we decide, after determining which hypothesis (null or alternate) the sample supports. A decision has two options. To reject the null hypothesis "reject H0" if the sample information supports it, or "don't reject H0" or "decline to reject H0" if it is not supported.

Hypothesis testing can be done for equality of the claim or less than/ greater than claim. Depending upon the purpose, equality or less than, greater than the test is decided. For equality testing we use two tailed test and for less than/ greater than test we use one tailed test. Following table provides comparison of two tail, left tail and right tail test.

TEST TYPE	TWO TAILED TEST	LEFT TAILED TEST	RIGHT TAILED TEST
NULL STATEMENT	H0: $\mu = \mu 0$,	H0: $\mu \geq \mu 0$,	H0: $\mu \leq \mu 0$,
ALTERNATIVE STATEMENT	Ha: $\mu \neq \mu 0$.	Ha: $\mu < \mu 0$	Ha: $\mu > \mu 0$.
OTHER POSSIBLE SCENARIOS OF ALTERNATIVE	Ha: $\mu < \mu 0$ (or) Ha: $\mu > \mu 0$.	NA	NA
EXAMPLE: CLAIM	Average person drinks 3 cups of coffee in a day H0: $\mu = 3$	Average person drinks more than 3 cups of coffee in a day H0: $\mu \geq 3$	Average person drinks less than 3 cups of coffee in a day. H0: $\mu \leq 3$
ALTERNATIVE: TO PROVE	Average person do not drink 3 cups of coffee in a day. Ha: $\mu \neq 3$	Average person drinks more than 3 cups of coffee in a day. Ha: $\mu < 3$	Average person drinks more than 3 cups of coffee in a day. Ha: $\mu > 3$
OTHER POSSIBLE ALTERNATIVES	Please note original claim is about equality. Hence alternative can also be set as greater than or less than Average person drinks more than 3 cups of coffee in a day Ha: $\mu > 3$ Average person drinks more than 3 cups of coffee in a day Ha: $\mu < 3$		

Do it yourself: Write down null and alternative hypothesis statement for following case and determine each case is one tailed (right or left) and two tailed test.

- To test (claim) if the mean number of hours spent working per week by college students who hold jobs is different from 20 hours.
- To test whether a bank's ATM is out of service for an average of more than 10 hours per month.
- To test if the mean length of experience of airport security guards is different from 3 years.
- To test if the mean credit card debt of college seniors is less than $1000.
- To test if the mean time a customer must wait on the phone to speak to a representative of a mail-order company about unsatisfactory service is more than 12 mail orders.

Exam Tip: To decide right or left tailed test, look at the cone of greater than or less than sign. Like if alternative is with > sign – cone is on right side hence it is right sided test.

6.3 Decision making process

In the process of hypothesis testing, our goal is to reject the null hypothesis, if our sample data supports this, else we fail to reject the null. For the decision making we have multiple approaches like t critical value method, confidence interval method and p value method. We will see all these methods one by one.

Hypothesis testing is conducted at a specific confidence interval. We can use any confidence interval for hypothesis testing, however some standard confidence intervals generally considered in hypothesis testing are 90%, 95%, 99% with one tail or two tailed test. We find t/z critical value using distribution table (z distribution, t distribution, etc) for given confidence interval. Null hypothesis is rejected or failed to reject at given confidence interval. Test statistics (t-stat) calculation does not require confidence interval. Following table provides the information about requirement of t stat or t critical value for given method.

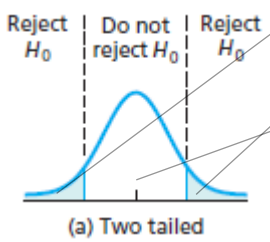
HYPOTHESIS TESTING METHOD	T-STAT CALCULATION	T/Z CRITICAL
T CRITICAL VALUE APPROACH	Required	Required
CONFIDENCE INTERVAL METHOD	Not required	Required
P VALUE METHOD	Required	Not required

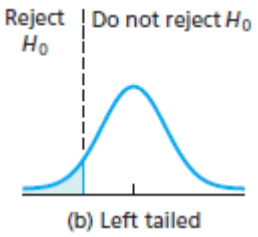
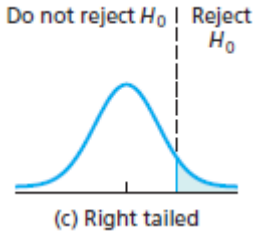
Irrespective of the method we choose, all these three methods result into same decision about the hypothesis statement.

6.3.a T critical value approach

This approach requires comparison of two values critical value (from distribution table) and test statistics (calculated value). Decision making is very simple using this approach –

Please note: For the sake of brevity, we will use t-stat for test statistics in following table. Test statistics can be z statistics or t statistics (calculated using same formula).

	Reject null	Fail to reject null
Two Tailed Test. 	| t-stat | > |critical value| i.e. (ignoring signs) if t-stat is greater than t critical	| t-stat | ≤ | critical value| i.e. (ignoring signs) if t-stat is less than t critical
Left Tailed Test	t stat < critical value	t stat ≥ critical value

	In left tailed test t critical is always negative	
Right Tailed Test 	t stat > critical value In right tailed test t critical is always positive.	T stat ≤ critical value

We already saw how to select one tailed and two tailed (right left) test in previous section. Now we will see how to get test statistics and critical values.

Test Statistics:

Assume, research suggests the average price of 1000Sq feet area house in a city is $50,000. We want to test this claim using hypothesis testing t critical method. We collected prices of 40 randomly selected houses (with 1000 sq feet area). Mean price using sample is $48000 and standard deviation is $2000.

Test statistics (z) $= \frac{\bar{x} - \mu 0}{SE}$

where,

SE is standard error of sample statistic (recall random sampling concept) = $\frac{\sigma}{\sqrt{n}}$

$\mu 0$ *is populaiton mean* (*population statsitic*)

X is sample mean (sample statistic)

σ *is standard deviation of sample set*

n is number of observations in sample set.

Applying this formula:

T stat (z) = $\frac{(48000 - 50000)}{\left(2000 — \sqrt{40}\right)}$ = - 2000 / 316 = -6.32

In the next step we will compare test statistics with critical value. For now, let's assume critical value range is -2.5 to 2.5 (we will see how to find critical value in next section). This is two tailed test and t stat (without sign) 6.32 > critical value 2.5, hence we reject the null statement. Which means true mean price of houses in a city is not equal to $50000.

What is exactly happening in test statistic calculation?

In the numerator we calculate the difference of sample mean vs population (hypothesized) mean. In the denominator we calculate standard error, i.e. error in sample mean. To lower the standard error the only option is increasing number of samples. Lower standard error increases absolute t stat. Higher t stat increases the chances of rejecting null hypothesis. Assume in the previous illustration we selected the sample of 5 only. t stat with n = 5 = -2.23. Hence absolute t stat is 2.23 < 2.5. We fail to reject the null. Our goal was to reject null, and we fail to reject null due to lower sample size.

Critical value:

Critical value is the rejection point for test statistics, which decides to reject or fail to reject the null statement. Critical value depends on choice of distribution and choice of confidence level. For mean testing, we must choose between standard normal distribution and students t distribution depending upon the situation. Confidence level of 90%, 95%, or 99% are often preferred choices in hypothesis testing.

When z distribution is the choice of distribution for hypothesis testing, we have standard critical values for most commonly used confidence intervals. Significance level is 1 – CL and denoted by α. For confidence level of 95% significance is 5%.

LEVEL OF SIGNIFICANCE (1-CL)	TWO TAILED CRITICAL VALUE	ONE TAILED CRITICAL VALUE
A = 10%	± 1.65	+ 1.28 or – 1.28
A = 5%	± 1.96	+1.65 or -1.65
A = 1%	± 2.58	+2.33 or -2.33

Above table provides standard values which you must remember for exam purpose. GARP mostly asks questions containing standard confidence level. But you should also learn to find critical values for other significance level just to be on the safer side (i.e. find z value from z distribution table).

How to make a choice between t distribution and z distribution?

In this chapter we use standard normal distribution to determine critical values in hypothesis testing about unknown parameters. With the support of CLT we can use normal distribution irrespective of actual population distribution is normal or not.

However, students t distribution is better choice when sample size n is small < 30 and population is not normally distributed. Remember CLT applies for larger sample size and hence for small sample size, standard normal distribution cannot be used if data is not normally distributed.

In simple words, use t statistics (students t distribution) for hypothesis tests of the population mean, if the population sampled has unknown variance and either of the following condition is satisfied,

- The sample is large enough ≥ 30 or
- The sample is small enough < 30 but the population is normally distributed or approximately normally distributed.

What effect does the choice of distribution have on hypothesis testing if the distribution is a student's t distribution (t test)?

Test statistics calculation is same except for some notation changes.

Test statistics Z stat = $\frac{\bar{x}-\mu}{SE}$

Test statistics t_{n-1} = $\frac{\bar{x}-\mu}{SE}$

Where SE is standard error of sample mean = $\frac{\sigma}{\sqrt{n}}$

We can see both z stat and t_{n-1} stat formulas are same.

Critical value for t test is found in students t distribution table for n-1 degrees of freedom. For two tailed hypotheses testing at 95% confidence level, with sample size of 10, can be found in t distribution at 5% α two tailed for df = 9 (10-1).

Following table provides one tailed t values. To find two tailed value in one tailed table use α /2 = 2.5%. Hence critical value df of 9 = 2.82144 for 5% significance level.

For one tailed hypothesis testing at 95% confidence level, with sample size of 10. Critical value (df = 9) = 2.262

DF	0.25	0.1	0.05	0.025	0.01	0.005
1	1.00000	3.07768	6.31375	12.70620	31.82052	63.65674
2	0.81650	1.88562	2.91999	4.30265	6.96456	9.92484
3	0.76489	1.63774	2.35336	3.18245	4.54070	5.84091
4	0.74070	1.53321	2.13185	2.77645	3.74695	4.60409
5	0.72669	1.47588	2.01505	2.57058	3.36493	4.03214
6	0.71756	1.43976	1.94318	2.44691	3.14267	3.70743
7	0.71114	1.41492	1.89458	2.36462	2.99795	3.49948
8	0.70639	1.39682	1.85955	2.30600	2.89646	3.35539
9	0.70272	1.38303	1.83311	2.26216	2.82144	3.24984

6.3.b Confidence interval method

Second method for hypothesis testing is confidence interval method. This method is simple and straight forward. This method does not require test statistics calculation.

Illustration:

Assume hypothesized mean value is equal to 50. Sample size of 50 is used to conduct hypothesis testing. Sample mean is equal to 48 and standard deviation is 3.5. At 90% confidence level should we reject the null (assume two tail).

Construct confidence interval using

Sample mean ± Critical Value X SE = 48± 1.65 X $(3.5/\sqrt{50})$ = 47.18 to 48.82.

We reject null if hypothesized mean is outside this range. And fail to reject null if hypothesized value falls within this range.

Following diagram shows rejection region for two tailed test, upper tail and lower tail test.

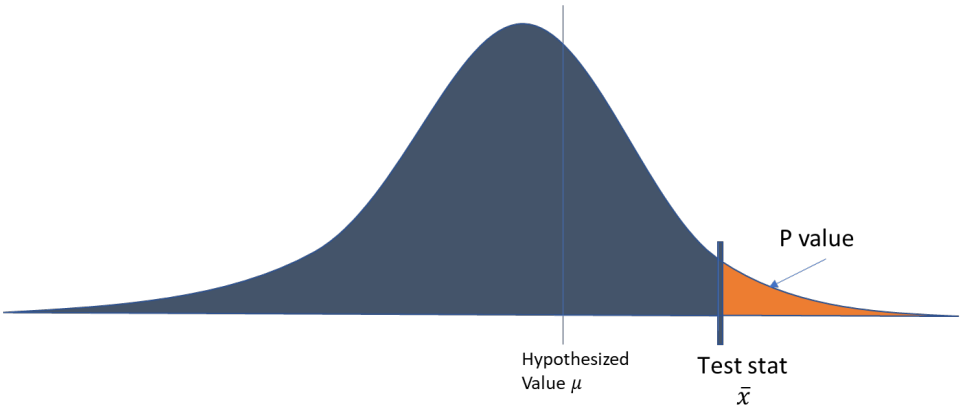
Based on research, the average value students debt on senior college student is $10,000. For hypothesis testing researcher took sample of 50 students. The average loan value of sample $9500 and standard deviation is $2000.

1. Conduct the hypothesis test at 95% confidence level to check if the original statement is wrong and average loan value is less than $10000. (Left tailed test)
2. Conduct the hypothesis test at 95% confidence level to check if the original statement is wrong and average loan value is more than $10000. (Right tailed test)

	LEFT (LOWER) TAILED TEST	RIGHT(UPPER) TAILED TEST
NULL HYPOTHESIS H_0	H0: $\mu \geq 10000$	H0: $\mu \leq 10000$
ALTERNATE HYPOTHESIS H_A	Ha: $\mu < 10000$	Ha: $\mu > 10000$
RULE: **TO REJECT NULL**	If population mean is < Sample mean + SE X critical value	If population mean is > sample mean – SE X critical value
STANDARD ERROR SE	$\frac{2000}{\sqrt{50}} = 282.84$	282.84
CRITICAL VALUE AT 95% ONE TAIL	1.65	1.65
CALCULATION	9500 + 282.84 x 1.65 = 9967	9500 – 282.84 x 1.65 = 9033.314
CONCLUSION:	10,000 < 9967 Reject null.	10000 < 9033.314 Fail to reject null.

P value method

P value is the lowest level of significance for which null can be rejected. P value is the probability calculated using test statistics. This method is simplest among all and most used in practice. Simply calculate test statistics (t stat or z stat). Using the distribution table, find out the probability area captured by test statistics. This is called as p value.



Decision rules using p value method

Reject null if p value ≤ α

Calculation of P value

We already know how to calculate test statistics (z score or t score). In the following example we will assume some test statistics and will find p value using standard normal distribution table.

Test statistics	P value
2.11 upper tail	1.74%
1.85 lower tail	3.216%
± 1.5 two tail	6.68% X 2 = 13.36%

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-2.90	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
-2.80	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
-2.70	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
-2.60	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
-2.50	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
-2.40	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
-2.30	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
-2.20	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
-2.10	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
-2.00	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
-1.90	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
-1.80	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
-1.70	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
-1.60	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
-1.50	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
-1.40	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
-1.30	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
-1.20	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
-1.10	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
-1.00	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786

6.4 Errors in hypothesis testing

We should be aware that no matter the design setup of our test, we are at risk of committing an error of making the wrong decision. Decision can be wrong in two ways; we may reject the null which is true in reality and we may fail to reject null which is not correct in reality. Say accused is standing in trail in front of judge and decision is pending. Judges don't know the truth about the accused's crime, and decision is to be made based on available evidence. There are two possible errors can happen in judgement, accused proven guilty but he didn't commit crime and accused is not proven guilty due to lack of strong evidence, but he committed crime. Similarly in hypothesis testing, decision is made using available procedure-based evidence, and two errors can be made by analyst. Two errors in hypothesis testing are,

Type I error: Occurs when H0 null hypothesis is true but wrongfully rejected the null hypothesis. The probability of Type I error is equal to significance level (alpha).

Type II error: Occurs when a false hypothesis is not rejected (fail to reject null). The probability of Type II error is beta.

		True Situation	
		H0 True	H0 False
Decision	Fail to reject H0	Correct Decision	Type II error (β error)
	Reject H0	Type I Error (α error)	Correct Decision (Power of Test) (1- β)

Before we begun the hypothesis testing, we need to make a choice between Type I error and Type II error. There is no way to reduce both the errors. These are mutually exclusive errors. Attempt to reduce Type I error will increase Type II error and vice versa.

Measuring probability of error:

Type I error: Type I error is we reject null hypothesis, but it is true. Type I error can be predetermined by alpha (level of significance). If null is true, then in ideal world t stat would have fallen into fail to reject region which is 1 – significance level. Hence, making the wrong decision is simply probability of significance level (area out of fail to reject region).

Type II error: Probability of Type II error (β) is difficult to predetermine and comes with some caveats. Please note, we either reject null hypothesis or fail to reject null hypothesis but there is no case of accepting null. Type II error is the error of failing to reject null when null is false. If null is false (in reality) then we don't know its true value itself. Without knowing the true value and decision fail to reject, we cannot predetermine(easily) the probability of Type II error. Type II error can be measured by opting alternative procedures which also comes with certain conditions and restrictions, which is not the part of FRM curriculum. For exam purpose, just remember Type II error is equal to beta.

Power of Test

The goal of hypothesis testing is to reject null when null is false in reality. Probability of achieving this goal is known as power of test i.e. probability of rejecting false null. Power of test is simply 1 – Type II error (1-β). To increase the power of test, we can take certain measures like, improving sampling procedure, choosing appropriate distribution and significance level. Please note, Probability of Type I error and power of test increases or decreases at the same time. For example, if we increase the significance level, it will increase the Type I error and power of test at the same time and will decrease the probability of Type II error. By keeping significance level constant (i.e., constant Type I error), the only way to increase power of test and decrease Type II error is increasing sample size.

6.5 Testing difference between two population means

Previously we discussed the testing of population of mean equal to certain value. In testing of difference of two population mean we equate two population means or say difference between two population means is zero. For example, we want to check if the salary of FRMs working in two different states of USA California $\mu(c)$ and Alaska $\mu(a)$ is equal. The null in hypothesis setup would be,

Null Hypothesis: H_0: $\mu(c) = \mu(a)$. We can modify this setup into difference of means by,

Null Hypothesis: H_0: $\mu(c) - \mu(a) = 0$

Similarly, we can also setup test for difference greater than 0 and less than zero.

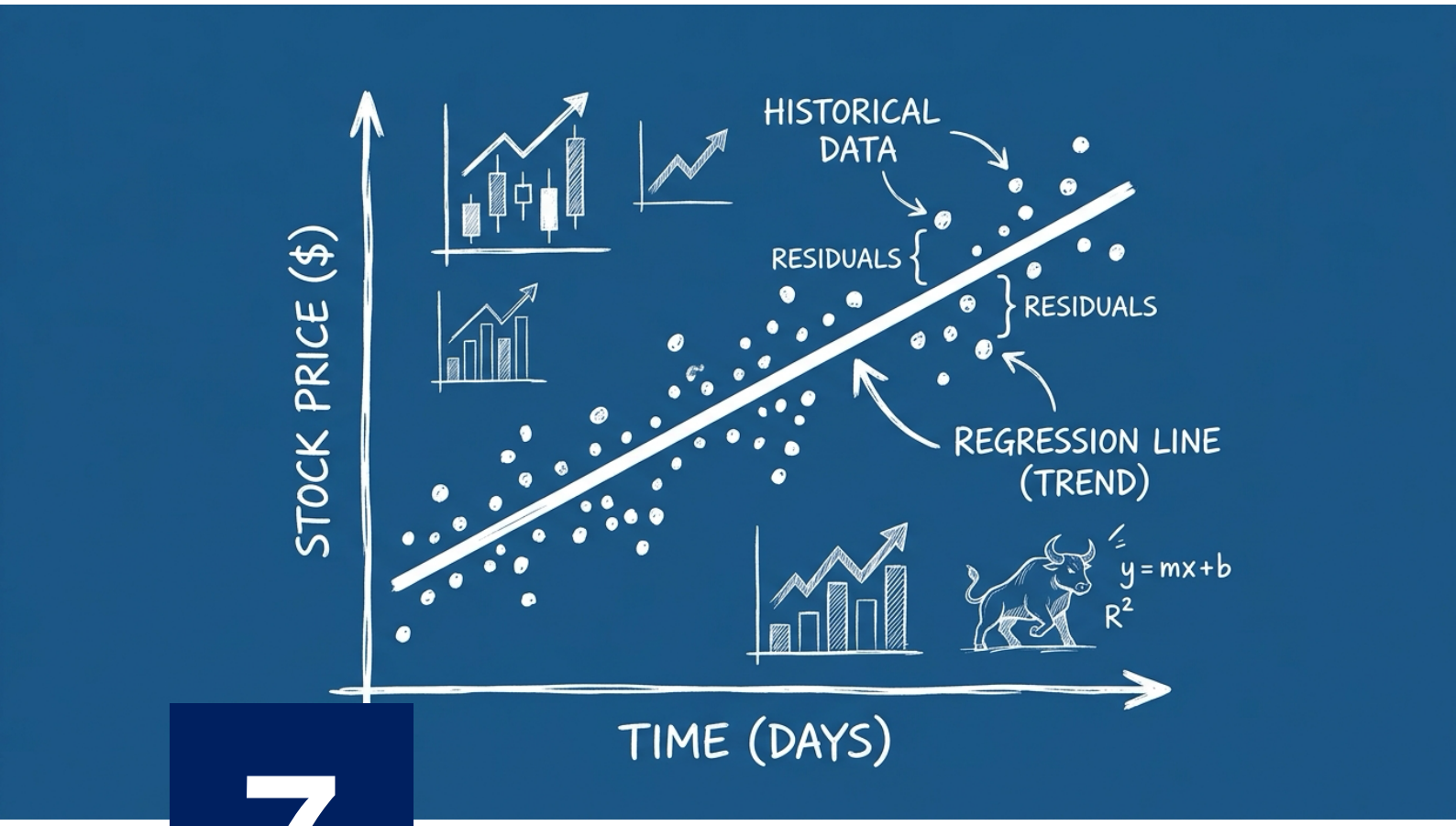
6.6 Multiple Hypothesis testing

Let's assume single hypothesis testing in which null hypothesis of smoking causes cancer in humans. In multiple hypothesis testing, multiple nulls are tested using same data set. Multiple hypothesis of above given single hypothesis can be,

- Smoking causes cancer in girls
- Smoking causes cancer in boys
- Smoking causes cancer in babies
- Smoking causes cancer in boys with long hairs

Problem with this testing is that it increases the probability of rejecting true null. *This increases probability of Type I error(alpha) exponentially.*

Reading 7 Linear Regression



7 Linear Regression

SCOPE OF THIS READING

This chapter introduces linear regression with a single explanatory variable and identifies the types of models that can be estimated using OLS. It explains the key assumptions underlying OLS estimation and characterizes the statistical properties and sampling distributions of the estimators. The chapter develops interpretation of regression output, including coefficient estimates and goodness-of-fit, and formalizes hypothesis testing and confidence interval construction for a single coefficient. It also clarifies the relationship among the t-statistic, p-value, and confidence intervals in statistical inference.

7.1 Introduction

Imagine you want to know how much studying affects your FRM exam score. You probably think that more studying means a higher score. But how can you measure this effect? You can use regression analysis to find out how much your score goes up for every hour you study. This is a way of finding a formula that links two things together. The thing you want to know, like your score, is called the response or dependent thing. The thing you use to guess it, like your study hours, is called the explanatory or predictor thing. In our example, your score is the response or dependent thing, and your study hours are the explanatory or predictor thing.

Example in Equation form

Exam score = a + b (number of hours)

In this equation 'a' is intercept. Intercept is the value which is taken by dependent variable if independent variable is 0. Assume a student appeared for exam without a single minute of preparation, she can still score say 20 in exam by just randomly ticking answers, which is intercept of 20. The b given in equation is known as correlation coefficient or slope coefficient which indicates change in dependent variable per unit change in independent value. Assume slope is 0.12 in our example. So, if a student prepares for 500 hours he is likely to score

Exam score = 20 + 0.12 X 500 = 80

This equation is a linear regression equation that helps us to estimate the dependent variable based on the independent variable. The main elements of a linear regression equation are the intercept, the slope, and the independent variable (explanatory variable). The independent variable is given, meaning we don't compute it, it is an input. The intercept and the slope are computed values using the past data. The aim of the regression equation is to find the parameters intercept and slope coefficient using the past data and then predict the dependent variable using these parameters. We will see how to find the intercept and the slope later in this chapter. First, let's look at some key terms and what they mean.

7.2 Steps in linear regression

Following are the steps used in regression analysis.

- **Stating the problem:** The problem statement is very important because il defined problem can result into inconclusive results. Example, what is the relationship of students score in exam and hours of preparation?
- **Variable selection:** Variable is selected using cause and effect analysis. We want variables which affect dependent variable. In our example, we selected variable hours of preparation. We can also select two or more variables, like number of hours in reading, number of questions solved, and number of mock test papers solved. When only one explanatory variable is used, we cover it into one variable regression (discussed in this chapter) and for two or more than two variables we use multiple variable regression (discussed in next chapter).
- **Data collection:** Samples data is collected using various methods. Here it is important to cover to collected data based on variable selection.
- **Model specification:** Model is specified in the form of equation. Model can be linear or non linear with one variable or multiple variables. We will discuss difference between linear and non linear regression variable below.

- **Model Fitting:** This is the process of identifying parameters of regression equation checking the fit of the model. Model fitting is checking how good the model is in establishing relation.
- **Checking assumptions:** Model is tested for assumptions. Regression works properly only when underlying assumptions are fulfilled. In this step we check for those assumptions in model.
- **Forecasting:** Forecasting is the process of predicting values of dependent variable using independent variable. Please note, ideally in forecasting, values of explanatory variable should be in the range of data used for regression modelling. Considering our previous example, students score and hours of preparation we got highest value for preparation hours of 600. Now if someone enters 10,000 hours of preparation in model, it will produce inaccurate results.

7.3 Linear vs non-linear regression equation.

Consider the following model

$y = \alpha + \beta_1 x + e$

Where y is dependent variable, α is intercept, β_1 is slope coefficient, x is independent variable and e is error term.

This equation is like equation used in previous example is called linear regression model. Please note, linear in linear regression does not describe the relationship between dependent and independent variable. It relates to parameters like β entering the equation linearly (multiplicatively). Models like $y = a + b X^2$ or $y = a + b \ln(x)$ both are linear models even though x not entering in the equation linearly. This is because these equations can be transformed into generic linear equation like we saw above by simply replacing $x1 = \ln(x)$ or $x1 = x^2$in equation. Here x1 is transformed variable.

Linear regression must satisfy three essential properties –

- Relationship between dependent and independent variable must be linear in the unknown coefficients. i.e. model must have a single unknown coefficient multiplied by a single explanatory variable.
- Error must be additive. i.e. variance of the error must not depend on observed data.
- All explanatory variables must be observable.

Explanatory variable can be continuous, discrete or functions or one or more variable like x = x1 + x2. In linear model, parameters should enter multiplicatively. Consider this regression model $y = \alpha + \beta X^{\lambda} + e$, where λ enters the equation in power of x and not multiplicatively and also this results into two parameters for x. This violates first condition of linear regression. However, if regression model is $y = \alpha + \beta X^2 + e$, where x has power of two does not violate the first property because 2 is known value and property restricts only unknown parameter, hence this model is considered linear.

Note: We used exam scores and hours of preparation as an example of linear model. However, this situation is related to learning curves and the relationship is nonlinear. This relationship will have an s shape because there will be no score improvement for the first say 80 hours, but then there will be a high increase for the optimal hours of preparation. After reaching the peak, the increase will drop again. For example, studying for 800 hours may help you get 90 in FRM exam. But what if you study for 1600 hours? Clearly, these extra hours will not make much difference.

7.4 Ordinary least squares method

Let's consider linear regression model with one variable,

$Y = \alpha + \beta_1 x_i + e$,

This equation is population regression model. In this equation y is value we get using this model. X is observed value. We are now left with parameters α and β and error term e. We can these parameters by solving linear regression model using various method. Most common method of parameter estimation is ordinary least squares (OLS), so that the sum of the squares of e i.e. error term is lowest. We can rewrite this equation as $e = y - \alpha + \beta_1 x_i$ to represent error term, where goal is to reduce sum of squares of error term in OLS. The main objective of this process is to find the regression fit line with the help of parameters. In the following section we will see all the steps required to estimate parameters and find the regression line.

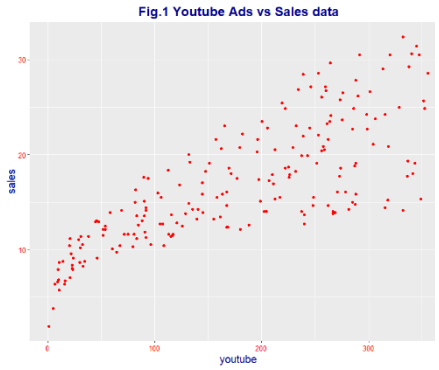
Note: For exam purpose you must understand steps given below to build robust concepts, but it is highly unlikely to get question which requires parameter estimation using data. If you get such question, TI BA II plus calculator provides function can be used which requires you to feed the data to get the answer. You can find free calculator course on our website in free course section

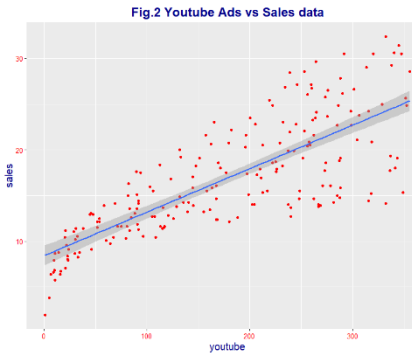
Case study: You are provided with the task to find the effect of YouTube marketing ads on daily sale of smartphone. Following table provides daily YouTube ads (as independent variable) and sales data (dependent variable). Table given here provides 5 random values (extract of full table). For actual analysis we will use 25 observations.

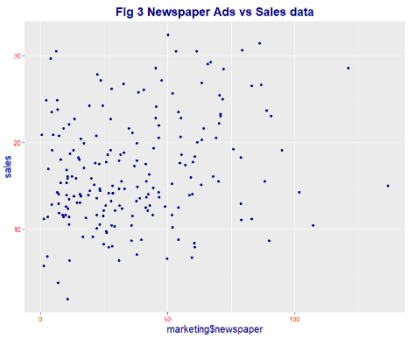
Daily sales (Y)	YouTube ads(X)
26.88	234.48
6.72	15.84
12.36	167.4
12.6	187.92
10.44	20.28

7.4.a Visualizing Data

Before we begin our regression analysis, it is good practice to start with visualizing data. Fig 1 shows positive relationship between YouTube ads and sales. Sales are increasing with increasing number of YouTube ads. Fig 2 is same graph but with regression line. Our goal in regression analysis is to get this line. Fig 3 provides comparison of newpaper ads and sales which shows very weak relation. Because fig 1 shows strong relationship between youtube ads and sales, hence it is wise to perform regression analysis to evaluate the impact of youtube ads on sales (Y).







Note: These graphs are produced using 200 observations to provide better visualization but we will use only 25 randomly selected observations in our regression analysis to ensure page space saving.

7.4.b Parameter estimation

In the parameter estimation we will focus on α and β calculations. Consider linear regression equation, Y = α + β X + e . When we use least squares method to estimate parameters α and β, we use least squares regression model $\hat{Y} = \hat{\alpha} + \hat{\beta}$ X . We use alpha hat and beta hat in the least square equation which signifies that $\hat{\alpha}$ *and* $\hat{\beta}$ are estimates of α and β because they are the solution of least square method.

Out of these two parameters, we first estimate β and with the help of β we will solve α.

7.4.c Concept and calculation of Beta (via correlation)

We already discussed correlation coefficient in basic statistics (reading 0). Correlation coefficient is the standardized version of covariance calculated using ratio of covariance of two variables and standard deviation of respective variable. Positive correlation shows that the movement of two variable is in positive direction. For example, we are performing analysis on RIL Inc stocks and market index and as per the analysis correlation between these two is say 0.60. This means both RIL and index are positively related i.e. increase in market index increases RIL. But this also means increase in RIL increases market. This is because correlation of X to Y and Y to X is one and the same thing. If we apply this to our case, correlation of RIL to index and vice versa is same. One more concern with correlation measure is that it only provides direction and strength of the directional movement between two variables. But it fails to provide the magnitude of impact on one variable by another. Using our case correlation fails to provide what is the increase in RIL if market index increases by 10%.

This problem is solved by beta measure. Beta provides the impact of x with respect to y. Please note that in regression analysis Y is variable of important because it is estimated and X is not because it is observed. Say beta of RIL with respect to market index is 1.5, which means when market increase by 10%, RIL stock will increase by 15% (1.5 of 10%). Unlike correlation which ranges from -1 to +1, beta has no range which can be any positive or negative value. Negative beta signifies negative relationship with variable. To calculate beta, we will use correlation only. In regression $\hat{\beta}$ is beta only.

$$\widehat{\beta x} = \frac{Cov\,(XY)}{Var(X)} = Correlation\,(XY)x\,\frac{SD(Y)}{SD\,(X)}$$

Using the following table, we will see how to reach at Beta of YouTube ads with respect to daily sales(25 observations – sample data).

Daily Sales	**YouTube Ads**					
(Y)	(X)	(Y - Yavg)	(X - Xavg)	(Y-Yavg)^2	(X-Xavg)^2	(Y-Yavg)(X-Xavg)
26.88	234.48	12.10	97.92	146.43	9589.27	1184.97
6.72	15.84	-8.06	-120.72	64.95	14572.16	972.87
12.36	167.4	-2.42	30.84	5.85	951.40	-74.62
12.6	187.92	-2.18	51.36	4.75	2638.34	-111.93
10.44	20.28	-4.34	-116.28	18.83	13519.92	504.54
13.68	115.44	-1.10	-21.12	1.21	445.85	23.21
22.8	205.56	8.02	69.00	64.33	4761.66	553.47
17.88	226.08	3.10	89.52	9.61	8014.69	277.60
12.48	125.52	-2.30	-11.04	5.29	121.78	25.37
20.52	212.4	5.74	75.84	32.96	5752.43	435.41
12.48	53.4	-2.30	-83.16	5.29	6914.79	191.19

17.52	93.84	2.74	-42.72	7.51	1824.59	-117.07
17.28	207	2.50	70.44	6.25	4962.47	176.17
23.04	232.44	8.26	95.88	68.24	9193.89	792.09
3.84	4.92	-10.94	-131.64	119.67	17327.83	1439.98
10.32	79.32	-4.46	-57.24	19.88	3275.87	255.22
15.48	248.28	0.70	111.72	0.49	12482.43	78.30
28.44	238.68	13.66	102.12	186.62	10429.47	1395.11
11.64	113.04	-3.14	-23.52	9.85	552.96	73.82
12.12	53.64	-2.66	-82.92	7.07	6874.93	220.49
8.04	22.44	-6.74	-114.12	45.42	13022.28	769.05
7.08	20.64	-7.70	-115.92	59.28	13436.33	892.45
11.64	71.52	-3.14	-65.04	9.85	4229.58	204.16
14.16	332.04	-0.62	195.48	0.38	38214.31	-121.04
20.04	131.76	5.26	-4.80	27.68	22.99	-25.23
Mean Y	Mean X	0.00	0.00	927.70	203132.23	10015.56
14.7792	136.5552	Sums				

Beta = $\widehat{\beta x} = \frac{Cov\,(XY)}{Var(X)} = \frac{\frac{10015}{24}}{\frac{203132}{24}}$ = 0.050 (approximated)

Beta of 0.05 tells us that for every one unit increase in independent variable, dependent variable will increase by 0.05.

We can also calculate beta using second formula which uses correlation. Answer will be same in both the cases. Please note we can solve beta using TI BA II Plus calculator.

7.4.d Intercept

Intercept is very simple to calculate and in meaning. Intercept is the value of dependent variable when independent variable is zero. Using our case, total sales even if we don't publish any ad on YouTube. To calculate intercept we use average of dependent and independent variables. Regression equation for intercept calculation

$\bar{Y} = \hat{\alpha} + \beta \bar{X}$

14.78 = α + 0.05 * 136.555

Therefore, $\alpha = 14.78 - 0.05 * 136.555 = 8.04$

Final equation using intercept and slope is

Sales = 8.04 + 0.05 x (YouTube Ads) – Linear regression equation.

7.4.e Error Term and Sum of squared errors

In this equation, sales is estimated using regression equation. If we use first observed value (from table above) of x (independent variable) we get $\hat{Y}$ = 8.04 + 0.05 x 238.48 = 19.96. But actual sales with YouTube ads of 238.48 is 26.88. This difference is error term e. The regression equation with error term is

$Y = \alpha + \beta X \pm e$

$Y = 8.04 + 0.05 * 238.48 + 7.27 = 26.88$ (you may find rounding off error)

Using similar methods, we will find errors for all the observed value to calculate sum of squared errors as given in following table.

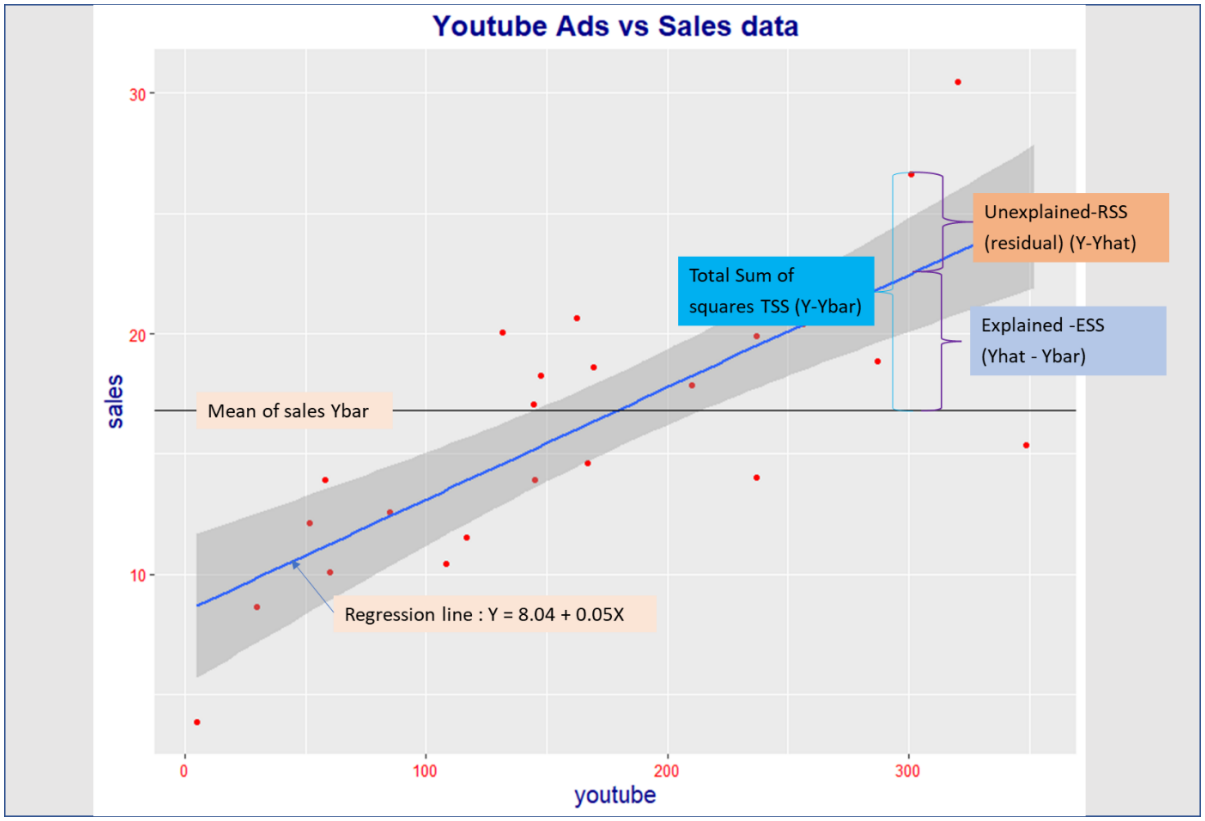
Explanatory X	Actual Y	Predicted $\hat{y}$	Residual e $Y - \hat{y}$	Residual squares $(Y- \hat{y})$ ^2	Explained Squares $(\hat{y} - \bar{y})$^2
234.48	26.88	19.61	7.27	52.89	23.31
15.84	6.72	8.83	-2.11	4.44	35.43
167.4	12.36	16.30	-3.94	15.52	2.31
187.92	12.6	17.31	-4.71	22.20	6.41
20.28	10.44	9.05	1.39	1.94	32.87
115.44	13.68	13.74	-0.06	0.00	1.08
205.56	22.8	18.18	4.62	21.33	11.58
226.08	17.88	19.19	-1.31	1.72	19.48
125.52	12.48	14.24	-1.76	3.08	0.30
212.4	20.52	18.52	2.00	4.00	13.98
53.4	12.48	10.68	1.80	3.24	16.81
93.84	17.52	12.67	4.85	23.49	4.44
207	17.28	18.25	-0.97	0.95	12.06
232.44	23.04	19.51	3.53	12.48	22.35
4.92	3.84	8.29	-4.45	19.79	42.12
79.32	10.32	11.96	-1.64	2.68	7.96
248.28	15.48	20.29	-4.81	23.12	30.35
238.68	28.44	19.81	8.63	74.40	25.35
113.04	11.64	13.62	-1.98	3.92	1.34
53.64	12.12	10.69	1.43	2.04	16.71
22.44	8.04	9.15	-1.11	1.24	31.66
20.64	7.08	9.06	-1.98	3.94	32.66
71.52	11.64	11.57	0.07	0.00	10.28
332.04	14.16	24.42	-10.26	105.22	92.90
131.76	20.04	14.54	5.50	30.22	0.06
Mean of Y	14.7792		Sums	433.87	493.82

Table provides residual sum of squares = $\sum e^2 = \sum(Y - \hat{y})$^2 = 433.87.

OLS regression is the process to reduce this sum of squared residuals.

7.5 R^2 - Explained vs unexplained variance in regression

The purpose of the regression analysis is not to determine the exact impact of the independent variable on dependent variable. Regression analysis is approximation of the impact of independent variable on independent variable. Because of this approximation we will get some error in actual value of dependent variable and predicted value of dependent variable as shown in the above table. Residual is that difference which is not explained by the regression line or equation. Now the question is how much portion of total difference is explained by the regression line? This graph shows explained and unexplained variance by regression line (YouTube ads and sales).



Total variation (underlying philosophy): Lets ignore regression equation line for now and just focus on the sales data and mean of sales ($\bar{y}$). Assume you are provided with the data of sales same as above and you are told to provide estimate of sales for given number of YouTube ads. With no other information available the best option available for you is to report average of sales which is 14.77. But if we compare average with actual data there is huge difference. This difference is total variance from mean to actual value of observation. If we take total of all such differences, we get Total sum of squares (TSS).

$$TSS = \sum(Y - \bar{Y})^2$$

To improve our estimate of sales for given YouTube ads we used regression analysis which provides linear regression line. This regression line provides the predicted Y. If we compare average of Y and predicted Y, predicted Y is closer to actual Y (ref above fig). This improvement in our estimate value comes from the predicted Y i.e. variance explained by predicted Y. Sum of all explained portion is called ESS (explained sum of squares).

$$ESS = \sum(\hat{Y} - \bar{Y})^2$$

Now if we compare predicted Y with actual Y we can see predicted Y still fails to give us the exact result of actual Y. This is the error of regression line. Sum of square of all errors is RSS (residual sum of squares.

$$RSS = \sum(Y - \hat{Y})^2$$

Hence, we can establish, total variation is sum of explained variation and unexplained variation.

TSS = ESS + RSS

7.5.a Measure of Fit R^2

Measure of fit in simple language, how well the regression line is able to explain the actual data. R^2 measures the fit of regression line. R^2 also called as coefficient of determination ranges from 0 to 1. R of 0 means regression line fails to explain variation and 1(or 100%) means regression equation perfectly explains the variation.

$R^2 = \frac{Explained\ variance}{Total\ variance} = \frac{ESS}{TSS}$ or $(1 - \frac{RSS}{TSS})$

R^2 is simply the ratio of portion of variance explained by regression line and total variance.

R^2 of previous example $= \frac{493.82}{927.69} = 53.23\%$ or 0.5323

R^2 of 0.5323 means regression line is able to explain 53.23% of the total variance. R^2 more closer to 1 is better.

Important note for Exam: R^2 is (due to mathematical setup) also square of correlation coefficient or we can say correlation coefficient is equal to under root of R^2. Please note this property is only applicable for linear regression with one variable. This is not applicable for multiple variable regression which we will study in next reading.

7.6 Dummy Variable

So far we have considered that explanatory variable is numeric. But what if the variable is categorical? Regression equation cannot handle any categorical information. Suppose in hours of preparation and exam scores case study, we also want to examine how working and non-working candidates affect exam score. This will lead to this equation.

Exam score $= \alpha + \beta 1$ (Hours of preparation) $+ \beta 2$ (work status) $+ e$

We cannot use work status as it is in the above equation, because it is qualitative information. To deal with this issue, we use a dummy variable or a binary variable. A dummy variable can only be 0 or 1. We replace work status with a dummy variable by assigning 1 to working and 0 to not working. Including a dummy variable in the equation affects the intercept and the slope of other variables. We rewrite the above equation using a dummy variable.

Exam score $= \alpha + \beta 1$ (Hours of preparation) $+ \beta 2$ (D) $+ e$, where D = 0,1

7.7 Properties of OLS estimators

The derivation of OLS estimators requires only one assumption that variance of explanatory variable X is positive. This property is easy to verify. However, to ensure viability of OLS estimators we need additional assumptions which are

- **Shocks are mean zero - E(e|X) = 0:** This property is known as mean independence which requires that X provides no information about the location of error. This also means correlation of X and error term is zero. This also implies that the unconditional mean of error is zero. E(error) = 0. Please note this assumption is not directly testable, as shocks are estimated and can not be evaluated before we perform regression. However, non-violation of this assumption can be established by careful review of data generating process for variables. Following are the examples of data generating process where this assumption is violated –

- **Sample selection bias or survivorship bias:** It occurs when some observations are not considered because of missing values of Y. One good example of survivorship bias is, when we (Falcon) collect data on FRM candidate results and their exam preparation process, we majorly get the data from candidates who cleared exam. Those who failed in exam don't share their results. This is survivorship bias where only successful candidates' data is part of data set. Survivorship words comes from vary interesting historical event from Word War II. Survivorship bias can be addressed using careful construction of dataset.
- **Simultaneity bias:** Simultaneity bias occurs when X and Y are simultaneously determined, and both X and Y are function of each other.
- **Omitted variable bias:** The model should not exclude important variables which are determinants of dependent variable. Omitted variable bias results into coefficients that are biased and may indicate relationship in model which is nonexistent in reality.
- **Attenuation bias:** Occurs when independent variables measured with error which results into inconsistent parameter estimation. Attenuation bias results into estimated slope which is flatter than true relationship.

- **Data are realization from iid random variables (all x and y are iid)**: It assumes that the pairs are iid draws from their joint distribution. The iid assumption affects the uncertainty of the OLS parameters estimators because it rules out correlation across observations. Note that OLS can be used in situation where variables are not iid by modifying method used to compute standard errors.
- **Variance of X is greater than 0.** Meaning variance of X is positive.
- **Variance of the error term (shock) is finite and constant:** Variance of e should not vary with X. This assumption is homoskedasticity (will be discussed in detail in next chapter).
- **No large outliers in x:** There should not be any outliers in x with high probability because OLS estimation is sensitive to large deviations. Simplest method to detect outliers is to visually examine data for extreme observations.

Implications of OLS assumptions: Assumptions imply that

- Estimators are unbiased so that E(α^) = α^ and E(β^) = β^ (when n is large enough).E(β^) = β^. (when n is large enough)
- Two estimators are jointly normally distributed. (and hence can be allowed to hypothesis test).
- Two estimators are jointly normally distributed (hence can be allowed to perform hypothesis test.

7.8 Properties of OLS estimators and their sampling distribution

Assuming the above is true, then the OLS estimators have a normal sampling distribution when the samples are large. The variance of this sampling distribution can be calculated from the data. Standard error is derived from the square root of the variance and can be used to do hypothesis testing with a t statistics and to make confidence intervals.

Variance of the slope estimators depends on two moments: the variance of shocks and the variance of explanatory variables. Furthermore, the variance of slope estimate increases with variance because accurately estimating slope is more difficult when the data are noisy. The variance of estimated slope is decreasing in variance of x.

The estimation error in intercept also depends on the variance of the residual and the variance of X. In addition, it depends on squared mean of X. If X has mean 0, then the asymptotic variance simplifies to variance and intercept. In practice, the CLT is used as an approximation so that slope is treated as a normal random variable that is centered at the true slope. The effect of sample size is clear in this approximation: t**he variance of slope decrease as the sample size increases.**

7.9 Hypothesis testing (all three methods)

We used sample data to arrive at OLS estimators and we know from the hypothesis testing chapter that this requires hypothesis testing to assure, estimators are representative of population parameters. We mainly deal with only one estimator in hypothesis testing of regression parameters which is slope coefficient $\hat{\beta}$. Intercept is not usually tested for very simple reason, it is not does not take the part in relationship of dependent and independent variable. Intercept is simply value taken by Y when X is zero. Hence in this part we will keep our focus on hypothesis testing of slope coefficient.

Hypothesis testing of regression coefficient is pretty similar to hypothesis testing of mean, which we say in hypothesis testing chapter. The change here is in calculation of standard error for slope coefficient which is more complicated than standard error of sampling distribution mean and hypothesis setup with fixed value. However, meaning and interpretation of both standard error of slope coefficient and standard error of sampling distribution mean is similar. The SE of slope is denoted as S_b.

Note: SE of β is difficult to calculate using pen and paper and hence it is highly unlikely to get question which requires calculation of SE in exam. GARP prefers testing candidates on application part here and most of the time values are given and we are asked to use these values in hypothesis test. In real life, we use statistics software, excel sheets or programming languages like R and Python which supplies us ready values. So our main job is to interpret these values. Similar approach is followed by GARP which is seen in recent exams.

Statistical Significance: Lets consider our previous regression model of YouTube and sales data.

Sale = 8.04 + 0.05 (YouTube Ads)

Here the slope coefficient of 0.05 shows the influence of ads on sales. The slope estimate is calculated using sample data, what if the real value of β is 0. It makes whole equation useless without the impact of X on Y. Hence our interest is in checking if the value of slope is not equal to zero i.e. significant. When we prove the same thing with the help of hypothesis testing we call it statistical significance.

Suppose we want to test the hypothesis that the value of the slope coefficient is equal to β0.

H0: β = β0 vs Ha: β ≠ β0

Even if it appears to be normal hypothesis setup, it is not. We are interested in testing the impact of X on Y. If slope coefficient is equal to 0 then there is no impact of X on Y. Hence, hypothesis is done with H0:β = 0 i.e. β0 = 0. It is not like we can not test null with certain value of β0 which is nonzero. But that's not the goal in hypothesis testing of regression coefficient.

We will perform hypothesis testing on slope coefficient using all three-methods, t stat, confidence interval and p value method. Following diagram summarizes all three methods which is enough for exam purpose.

Hypothesis testing of β (slope)

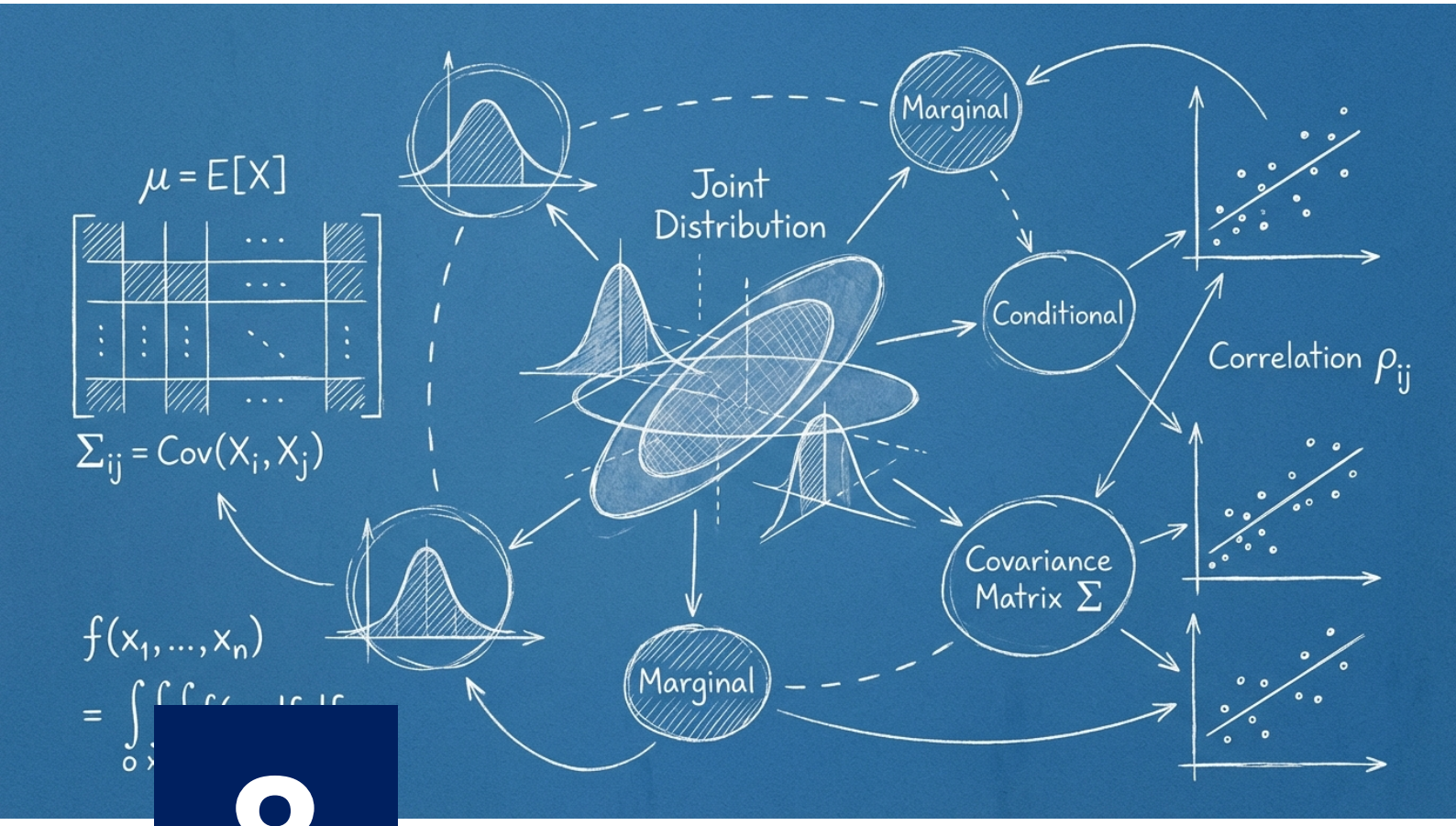
Suppose we want to test the hypothesis that the value of the slope coefficient is equal to β0.
H0: β = β0 vs Ha: β ≠ β0
Case study: Regression model Y = 0.20 + 0.65 X using 25 observations has Sb = 0.2. Determine if the slope coefficient is statistically significant at 5% significance level. (t

T stat method	Confidence interval	P value method
T stat calculation: $\frac{\beta - \beta 0}{Sb}$. Because we are testing for β0 = 0. This equation is simplified into T stat = $\frac{\beta}{Sb}$ = 0.65 / 0.2= 3.25	Confidence interval of β = β ± (tc X Sb) = 0.65 ± (2.07*0.2) = 0.236 < β < 1.06	P value is the smallest level of significance at which null can be rejected. P value will be provided in exam (because it is tedious to calculate by hand). P value calculated using t stat of 3.25 with df =23 at 5% significance is 0.00353
Reject Null If T stat > t critical	If range does not cover 0.	If P value < significance level
Null rejected because T stat 3.23 > t critical 2.07 Hence slope is statistically significant	Null rejected because 0 is not in the range of 0.236 < 1.06 Hence slope is statistically significant	Null rejected because P value 0.00352 < 0.05 Hence slope is statistically significant

Following is the regression analysis output produced using excel (with all important values highlighted). This provides the same information we used in the previous section. This table is presented just to show the exam style presentation of information. Also in real life we use these tables for regression.

REGRESSION STATISTICS				
MULTIPLE R	0.729596909			
R SQUARE	0.53231165			
ADJUSTED R SQUARE	0.511977373			
STANDARD ERROR	4.343274619			
OBSERVATIONS	25			
ANOVA				
	df	SS	MS	F
REGRESSION	1	493.8235925	493.8235925	26.17804769
RESIDUAL	23	433.8727915	18.86403441	
TOTAL	24	927.696384		
	Coefficients	Standard Error	t Stat	P-value
INTERCEPT	8.046260063	1.576787766	5.10294425	3.61481E-05
X VARIABLE 1	0.049305628	0.00963669	5.116448738	3.49609E-05

Reading 8 Regression with Multiple Explanatory Variables



8 Regression with Multiple Explanatory Variable

SCOPE OF THIS READING

This chapter extends regression analysis from single-variable to multiple regression frameworks. It distinguishes the underlying assumptions of single and multiple regression models and explains how to interpret coefficients in a multivariate setting. The chapter evaluates goodness-of-fit measures, including R^2 and adjusted R^2, in both contexts. It also develops joint hypothesis testing and confidence interval construction for multiple coefficients, enabling statistical inference on combined parameter restrictions.

8.1 Introduction

A regression model that uses more than one explanatory variable is a multiple regression model. In the previous reading we used YouTube ads and sales data in a regression model. In this model we only used one variable to explain the dependent variable. Suppose, we want to find out how Newspaper ads(X1) and YouTube ads(X2) affect sales data(Y). A regression model that could capture this relationship is given by

$y = \beta 0 + \beta 1x1 + \beta 2x2 + e$

In multiple linear regression we use more than one variable which creates the difficulty in graphing of this relationship. If we use only two variables in model then we can use surface graphs instead of line to represent relationship. For more than two variables, graphs are not produced hence graphs will not be used to explain concepts.

8.1.a Assumptions in linear regression with multiple regressor

Multiple linear regression model uses same 5 assumptions of one regressor model which we discussed in previous reading with some minor tweak and one additional assumption relating to collinearity.

Assumptions in multiple regressor model

Set 1: Only applicable for multiple regressor model

- Explanatory variables are not perfectly linearly correlated i.e. each variable must have variation that cannot be explained perfectly by the other variables used in the model.

Set 2: Applicable for both one regressor model and multiple regressor model (with some tweak)

- All the variables must have positive variances i.e. $\sigma^2 > 0$
- The error term e is assumed to have mean zero conditional on the explanatory variables.
- Random variables are assumed to be iid.
- No outliers in any of the explanatory variables.
- Constant variance for all explanatory variables (i.e. homoskedasticity – explained in next reading)

Beta coefficients in multiple regressor model are difficult to calculate by hand. In real life we use software packages (excel, R etc) to find out coefficients. Hence, coefficient determination is not tested in exam. GARP focuses on testing knowledge of assumptions and some key properties of multiple regressor models.

8.2 Interpretation of regression coefficients (Partial regresison coefficients

Regression model with one variable represents line, while the regression model with multiple variable represents surface (plane) for two regressors and hyperplane for more than two regressors. In multiple regression α (intercept) is called as constant coefficient, i.e. value of dependent variable when all the explanatory variables are zero (X1 = X2 … Xn = 0).

The slope coefficient β_i in multiple variable regression has multiple interpretations.

- Slope coefficient β_i is the change in dependent variable (Y) corresponding to a unit change in X_i, by keeping all the other explanatory variables constant. Where magnitude of change is not dependent on the value at which remaining explanatory variables held constant.
- The β_i is also called as the partial regression coefficient because it represents the change in Y contributed by Xi after adjusting for the other explanatory variables.

8.2.a Partial regression coefficients

To understand the partial regression coefficients, first we need to get the general understanding of regression modeling with multiple linear regression. Consider the following model,

$Y = \alpha + \beta 1X1 + \beta 2X2 + e$

To arrive at the slope coefficients, we go through following steps to ensure that the slope coefficients are calculated after keeping other variables constant (steps are not very important for exam purpose)

- **Step 1:** First fit the simple regression model using Y dependent variable and X1 explanatory variable to find the residual from this model denoted by e_{yx1}. Residual here is part of Y which is not linearly related to X1.
- **Step 2:** Fit the regression model using X2 as dependent variable and X1 explanatory variable to find the residual from this model denoted by e_{x2x1}. Residual is the part of X2 that is not linearly related to X1.
- **Step 3:** Fit the simple regression model that relates to residuals e_{yx1} (dependent variable) and e_{x2x1} (independent variable). We find the linear relationship between the Y residual and X2 residual.

Same procedure can be repeated for obtaining coefficients of X1 with slight modification. Here, the resulting regression coefficient represents the effect of X2 on Y after taking out the impact of X1 from both Y and X2. The slope coefficient βj is the partial regression coefficient because it represents the contribution of Xj to Y after both variables have been linearly adjusted for the other predictor variable. Slope coefficients in multiple linear regression are called partial coefficients because it partially (after certain restrictions) explains the impact of independent variable on dependent variable. On the other hand, in case of linear regression with one regression, slope coefficient supplies the impact of independent variable on dependent variable without such restrictions.

Following table provides regression coefficients, standard error, t stat and p value (generated using excel Data analysis tool pack)

	COEFFICIENTS	STANDARD ERROR	T STAT	P-VALUE
INTERCEPT	3.755840776	1.07519226	3.493180629	0.002166913
YOUTUBE	0.045732735	0.004639501	9.857252494	2.49062E-09
FACEBOOK	0.187792669	0.028348016	6.624543563	1.47147E-06
NEWSPAPER	-0.003606517	0.024472135	-0.147372386	0.884243856

Regression model using above results –

Sales = 3.76 + 0.046 (YouTube Ads) + 0.188 (Facebook Ads) + (-0.0036) (Newspaper ads)

Interpretations:

- Intercept: Sale value is 3.76, when all the independent variables is equal to zero. In simple language, even if we don't put any ads on YouTube, Facebook and newspaper, we can still get the sale of 3.76.

- Slope: Partial slope coefficient of Facebook ads is 0.1877, which shows the change in sale value per unit change in Facebook ads by keeping all the other value i.e. YouTube ads and Facebook ads constant. Please note, we are keeping other variables constant and not zero.

Note: Excel data analysis Tool pack also provides details on measure of fit which we will discuss below. Please note this section is not relevant for exam.

Process to perform regression analysis in MS Excel using data analysis tool pack?
MS Excel is a popular spreadsheet software that can perform various data analysis tasks, including regression analysis. In this article, we will explain how to perform regression analysis in MS Excel using the data analysis toolpack add-in. The data analysis toolpack is a collection of tools that can perform various statistical and engineering analyses on your data. It is not installed by default in MS Excel, so you need to enable it first before using it.

To enable the data analysis toolpack, follow these steps:

Click on the File tab and select Options.
Click on Add-Ins on the left sidebar.
In the Manage drop-down list, select Excel Add-ins and click Go.
In the Add-Ins dialog box, check the box next to Analysis ToolPak and click OK.
You may need to restart MS Excel for the changes to take effect.
Once you have enabled the data analysis toolpack, you can use it to perform regression analysis on your data.

To perform regression analysis

Arrange your data in a worksheet, with the response variable in one column and the explanatory variables in adjacent columns. Make sure to label the columns with descriptive names.
Click on the Data tab and select Data Analysis in the Analysis group.
In the Data Analysis dialog box, select Regression and click OK.
In the Regression dialog box, specify the following options:
Y Range: Select the range of cells that contain the response variable.
X Range: Select the range of cells that contain the explanatory variables.
Labels: Check this box if your data has labels in the first row.
Constant is Zero: Check this box if you want to force the intercept term to be zero in the regression equation. Otherwise, leave it unchecked.
Confidence Level: Specify the confidence level for the regression coefficients. The default is 95%.
Output Range: Select the range of cells where you want to display the output of the regression analysis. Alternatively, you can select New Worksheet Ply or New Workbook to create a new sheet or workbook for the output.
Residuals: Check the boxes for the residual plots and statistics that you want to include in the output. Residuals are the differences between the observed and predicted values of the response variable.
Click OK to perform the regression analysis and display the output.

8.3 Goodness of fit measures for single and multiple regressions (R^2 and adjusted R^2)

Three most used summary statistics in multiple regression are the R^2 , adjusted R^2 and SER (standard error of regression). In the previous reading we discussed the concept of R^2 and e (error in regression model). In this section we will extend these two concepts to measure the fit of multiple regressor models. Following is the result produced using excel data analysis toolpack

which provides the measure of fits which we will discuss one by one in the following sections. (same case study of sales and various modes of ads is used to produce this result).

REGRESSION STATISTICS	
MULTIPLE R	0.949705626
R SQUARE	0.901940776
ADJUSTED R SQUARE	0.887932316
STANDARD ERROR	2.08131354
OBSERVATIONS	25

8.3.a Standard error of regression

The SER (standard error of regression) estimates the standard deviation of error term e. Please note this is different from RSS (residual sum of squares). SER is a measure of the sprad of the distribution of Y around the regression line.

SER = $\sqrt{\frac{RSS}{n-k-1}}$. where n is total observations and k is total explanatory variables used in regression model.

SER calculation is same for both one regressor model and multiple regressor model. SER for one regressor uses n – 2 in denominator which is same as n – k – 1 (k = 1 for one regressor regression). For multiple regressor model used in our case study given above uses 5 observations and 3 explanatory variables hence for this model n – k – 1 = 25 – 3 – 1 = 21. Hence degrees of freedom for this model is 21.

8.3.b Adjusted R^2

In the previous reading we discussed about coefficient of determination R^2 also known as explanatory power of regression model. Lets assume we are working on regression model to regressor Y and we identified 10 variables which we thing are of importance and should be included in regression model. To produce a sound regressor model, we start with regression model by taking only one regressor which we think as most important and has highest impact on Y. Please note, in real life we will create 10 different one regressor models by separately using each identified variable as regressor. Then the model with highest explanatory power (R^2)is considered as first model. Second variable will be added on this one regressor model and then third variable and so on. Adding new variable in any existing model one regressor or multi is not free lunch. Following are the implications of adding new variable in regression model-

- With every added variable explanatory power will either increase or stays constant. Increase in R^2 means models explanatory power is increased and staying constant means it is not adding any explanatory power.
- Adding more variable means more control variables which we need to observe (and collect data of) which increases model complications.
- Most importantly, adding variable which increases R2 does not mean that it actually improves the fit of the model. If we keep adding variables it gives false estimates of regression fit of the data which needs correction.

The adjusted R2 is modified version of R2, which corrects inflated impact on R2 added by new variable introduced in the model. Unlike R2 the adjusted R2 may not necessarily increase with added variable. The increase in adjusted R2 depends on the explanatory power brought by new variable into the model. This measure works by considering the additional explanatory power brought by new variable and penalizing model for added variable. Following is the formula for adjusted R2.

Adjusted R2 = $1 - \left[\frac{n-1}{n-k-1}(1-R^2)\right]$

We can see in the formula, R2 is adjusted for added variables, hence this measure is called adjusted R2.

Following table provides some dummy models with increasing number of new variables and their respective R2 and adjusted R2.

MODEL	R^2	ADJ R^2 (N=25)	EXPLANATION / NOTE
Y =A + B1X1	0.44	0.42	Irrespective of the value of R2, adj R2 will always be lower than R2. We don't need compare R2 with adjusted R2. The thing to look here is if the adj R2 is increasing or decreasing with every added variable.
Y= A+ B1X1 + B2X2	0.56	0.52	Here adj R2 is increased to 0.52, hence we will assume X2 brings explanatory power into model and is worthy of adding)
Y= A+ B1X1 + B2X2+ B3X3	0.57	0.51	Adj R2 is decreasing after adding X3, even though R2 is increasing. This is what we look for in adj R2. X3 should not be added into model because it decreases adj R2. Reason for decrease in adj R2 is, very low explanatory power(R2) of only 0.01 brought by X3.
Y= A+ B1X1 + B2X2+ B3X3 + B4X4	0.68	0.62	X4 ads more explanatory power into the model and can be added but we need to test the explanatory power added by X4 after removing X3 from the model.

Question: Can R2 and adj R2 be negative in value (exam important).

Answer: R2 is explanatory power of any variable added. Theoretically the worst thing that can happen in regression model is, the first variable selected by us adds no explanatory power at all, which il result into R2 = 0. Hence R2 cannot be negative in any case. But for adj R2 it is different. If R2 is very low or added power by new variable is very less, in such cases adj R2 can be negative. The negative adj R2 is more prominent problem in case of low number of observations used for regression model. This happens because of mathematical formulation of adj R2. Let's assume R2 of 0.05 (5%) for one regressor model based on 5 observations. The adj R2 of this model is -0.27.

8.4 Joint hypothesis testing and confidence intervals for multiple coefficients in a regression

The fundamentals of hypothesis testing of regression coefficients are same as we discussed in the previous reading. For multiple regressor model just like one regressor model hypothesis testing of intercept is futile exercise. The focus is on testing of slope coefficients of explanatory variables.

Let's consider regression model from our case study

Sales = 3.76 + 0.046 (YouTube Ads) + 0.188 (Facebook Ads) + (-0.0036) (Newspaper ads)

Assume you are the analyst who produced this result in front of your CEO. Your CEO complained that YouTube ads adds no value in sales and asked you opinion on closing YouTube ad campaign. Obviously, you will go for hypothesis testing of slope coefficient(b1) of YouTube ads (X1) to test the statistical significance of b1. Hypothesis statement can be written as

H0: βj = βj,0 vs Ha: βj ≠ βj,0 (two sided test)

8.4.a Hypothesis testing for a single coefficient

In the above example of hypothesis testing, we are testing single regression coefficient. Hypothesis testing of single coefficient is same in both one regressor model and multiple regressors model which one modification. For one regressor, the degrees of freedom used is n – 2 and for multiple regressor model the degrees of freedom is n-k-1. This not actually difference but mentioned separately to avoid any confusions in the exam. We can use all three methods in same manner which we say in previous reading. The summary of hypothesis testing is given below.

This process is applicable only when we are evaluating the statistical significance of one slope coefficient. What if the your CEOs objection is, YouTube and Facebook ads both does not contribute to sales. In this situation, we have to test two or more slope coefficient. One approach is to test one coefficient at a time just like we did in the above example. Second approach is testing both the slope coefficients simultaneously.

Hypothesis testing of β (slope)

Suppose we want to test the hypothesis that the value of the slope coefficient is equal to β0.

H0: β = β0 vs Ha: β ≠ β0

Case study: Regression model Y = 0.20 + 0.65 X_1 + 0.25X_2 using 25 observations has Sb = 0.2. Determine if the slope coefficient is statistically significant at 5% significance level. (t critical = 2.07 at df = n-2 = 25-2 = 23)

T stat method	Confidence interval	P value method
T stat calculation: $\frac{\beta - \beta 0}{Sb}$. Because we are testing for β0 = 0. This equation is simplified into T stat = $\frac{\beta}{Sb}$ = 0.65 / 0.2= 3.25	Confidence interval of β = β ± (tc X Sb) = 0.65 ± (2.07*0.2) = 0.236 < β < 1.06	P value is the smallest level of significance at which null can be rejected. P value will be provided in exam (because it is tedious to calculate by hand). P value calculated using t stat of 3.25 with df =23 at 5% significance is 0.00353
If T stat > t critical	If range does not cover 0.	If P value < significance level
Null rejected because T stat 3.23 > t critical 2.07 Hence slope is statistically significant	Null rejected because 0 is not in the range of 0.236 < 1.06 Hence slope is statistically significant	Null rejected because P value 0.00352 < 0.05 Hence slope is statistically significant

8.4.b Joint Hypothesis Testing of two (or more slope coefficients) simultaneously

Hypothesis testing statement for testing of two slope is

H0: β1 = 0 and β2 = 0 vs Ha: β1≠0 and or β≠ 0

Putting this in the context of our case study of YouTube ads and Facebook ads impact on sales, we are testing for neither YouTube ads nor Facebook ads contribute to sales. In simple terms, slope coefficient of YouTube ads and or Facebook ads is not statistically significant (zero). The hypothesis is both the slope coefficients are zero is an example of a joint hypothesis testing of multiple regressor model. In this null hypothesis in above setup imposes two restrictions on the multiple regression model. Total restrictions in this case is 2. We will reject the null even if single slope coefficient is not equal to zero.

Assume a regression model with 10 variables and we want to test the statistical significance of 1st, 3rd, 4th and 5th variable. So the total number of restrictions in this hypothesis testing is = k = 4. If any one of the equalities under the null is false then the joint hypothesis itself is false. This gives the alternative hypothesis that at least one of the equalities in the null hypothesis does not hold.

F distribution for joint hypothesis testing:

One can think of using t statistics method or any method given above and use it on individual coefficients. This is appropriate in hypothesis testing of single coefficient but when it comes to testing of multiple coefficients simultaneously this method is unreliable. Because the equation involves two random variables, answering it requires the joint sampling distribution.

The F statistics is used to test joint hypothesis testing instead of t statistics (used for univariate distribution).

$F = \frac{\frac{(TSS - RSSu)}{k}}{\frac{RSSu}{n-k-1}}$, F test is always one tailed test.

This is generic F test used when all the regression coefficients in the model are tested at once. F test is then compared with critical F value with k degrees of freedom in numerator and n – k – 1 degrees of freedom in denominator. F value is mostly provided by GARP in exam. Hence, we are not concerned with F values.

Decision Rule: If F statistics > Critical F value – Decision: Reject Null

Note: In exam you can expect application of this decision rule where F statistics is either required to calculate or directly given and compare this value with critical F value.

Above F statistics is applicable when we are testing for all the regression coefficients. But in our case we are testing only two coefficients out of three. In this case we use different formula for F statistics. The model with all the variable is full model or unrestricted model and model with variable which we are testing with restriction is called partial model or restricted model.

Implementing an F test requires estimating two models. The first model to be tested is called full model and RSS is denoted by RSS_u. The second model is restricted model which imposes the null hypothesis on the unrestricted model and its RSS is denoted by RSS_R.

$$F\ stat = \frac{\frac{(RSS_R - RSS_u)}{q}}{\frac{RSS_u}{(n-k-1)}}$$

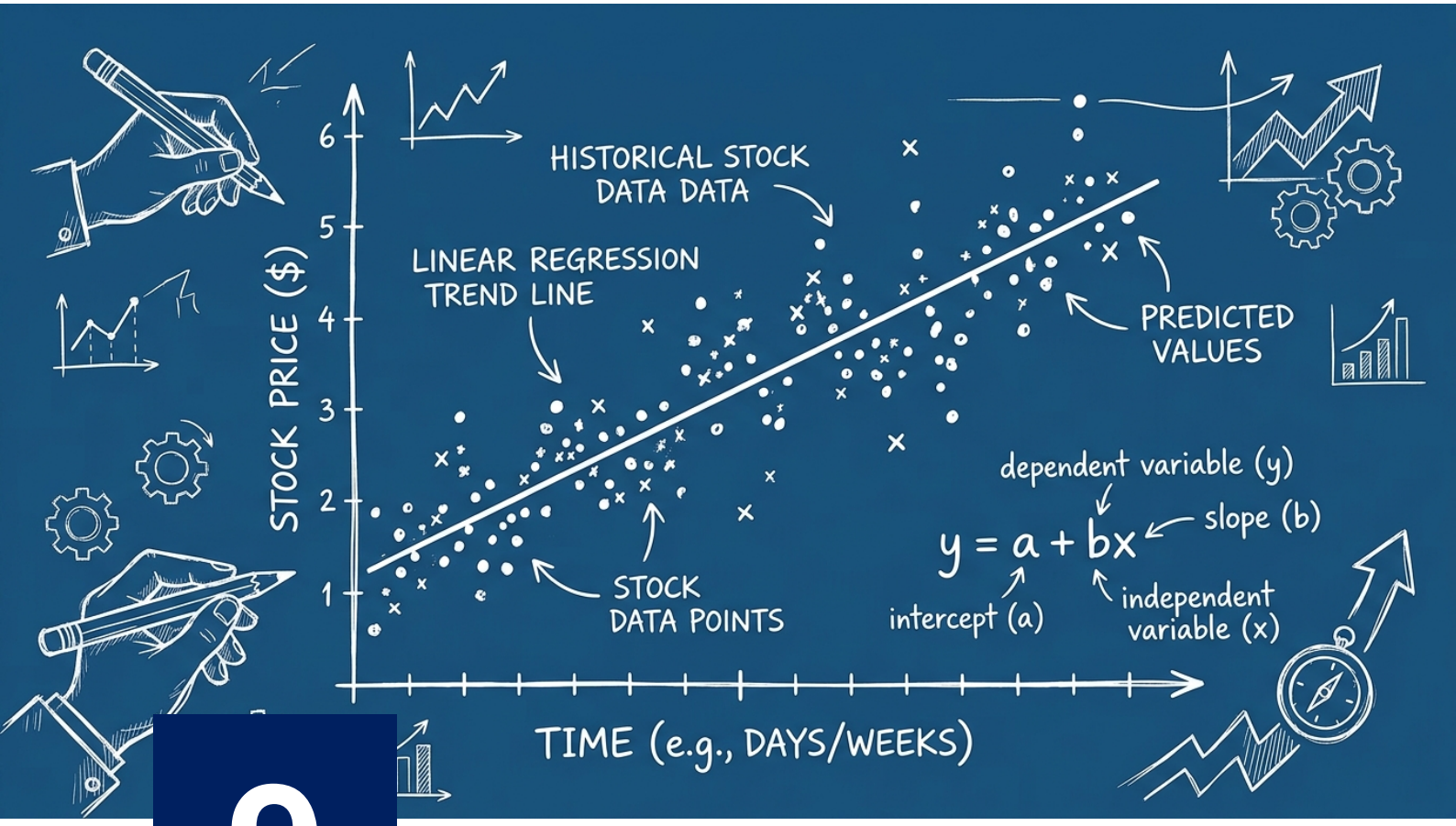
Where q is the number of restrictions imposed on the unrested model to produce the restricted model and k is the number of explanatory variables in the unrestricted model. F stat is then

compared to critical F value with q degrees of freedom in numerator and n – k – 1 degrees of freedom in denominator.

If the restriction imposed by the null hypothesis does not meaningfully alter the fit of the model, then the two RSS measures are similar, and the test statistic is small. On the other hand, if the unrestricted model fits the data significantly better than the restricted model, then the RSS from the two models should differ by a large amount so that the value of the F-test statistic is large. A large test statistic indicates that the unrestricted model provides a superior fit and so the null hypothesis is rejected.

Decision Rule: If F stat is greater than critical F value, Null is rejected which means unrestricted model appears to be adequate.

Reading 9 Regression Diagnostics



9 Regression Diagnostics

SCOPE OF THIS READING

This chapter examines key econometric issues affecting regression validity and inference. It explains how to detect heteroskedasticity and outlines approaches for handling heteroskedastic data. The chapter characterizes multicollinearity and distinguishes it from perfect collinearity, analysing their implications for estimation precision. It evaluates the consequences of omitting relevant variables versus including irrelevant regressors, and discusses model selection procedures in the context of the bias–variance trade-off. It also describes methods for identifying outliers and their influence on results, and specifies the conditions under which OLS satisfies the Gauss–Markov assumptions and is the best linear unbiased estimator (BLUE).

9.1 Why do we need regression diagnostics

In an ideal world, a model should include all variable that explains the dependent variable and exclude all the that do not, and the regression model should also adhere to the underlying assumptions of linear regression. In real life things are more complex. Choice of variables in regression model needs more consideration. Increasing variables in model makes it complicated and reducing variables in model makes it weaker in explanatory power. Also, assumptions of linear regression in model are not easy to test.

The concern with both the variable consideration and assumptions can only be reasonably tested after the regression model is formed and we get all the required estimates and test statistics. Hence, we conduct regression diagnostics. Once the model is produced, we diagnose it to verify that model fulfil required assumptions and also considers relevant variables. In this chapter we will see how to detect if model fulfil assumptions and how to correct these inaccuracies if required.

Following table provides the assumptions used in linear regression and concepts which we will study in this chapter which relates to these assumptions

Assumption / requirement of sound model	**Related concept (explained in this reading)**
Explanatory variables are not perfectly linearly correlated	Multicollinearity
Constant variance for all explanatory variables	Heteroskedasticity
No outliers in any of the explanatory variables.	Cook's measure
The error term e is assumed to have mean zero conditional on the explanatory variables.	Omitted variables

Following assumptions are not tested and assumed prior construction of the model

- All the variables must have positive variances i.e. $\sigma^2 > 0$
- Random variables are assumed to be iid.

9.2 Omitted Variable bias and extraneous variable and bias varianc tradeoff

An omitted variable is one which is related to the dependent variable but is not included in a model. Omitting a variable has two effects.

- The included variables absorb the effects of the omitted variable and changes in regression coefficients on the included variables. This results into, variables do not consistently estimate the effect of a dependent variable.
- The estimated residuals are larger than the actual residuals because residual now carry any effect of the omitted variable that is not captured by the included variables.

If the variable is correlated with the variable that has been omitted from the model and it determines the part of dependent variable, then OLS estimators will have omitted variable bias. Omitted variable bias occurs when both the following conditions are satisfied –

- Omitted variable is correlated with the included variables
- Omitted variable is determinant of the dependent variable

Omitted variable bias violates the OLS regression assumption "The error term e is assumed to have mean zero conditional on the explanatory variables." Reason for violation is, when the relevant term is omitted from the model, error term will carry its impact and hence error term also becomes the determinant of dependent variable (because it carries effect).

Bias due to omitted variable depends on the true coefficient of the excluded variable and correlation between included and omitted variable. If correlation between the included and omitted variable is high it results into higher bias. This is highly important for financial data, because financial data are generally correlated, so omitting variable creates bias and inconsistent estimates of included variable.

9.2.a Extraneous variable

An extraneous variable is the one that is part of the model, but it is not required to be included. In simple terms we can say, if included variable is irrelevant to the model then it is extraneous variable. This means the true slope coefficient of this variable is zero.

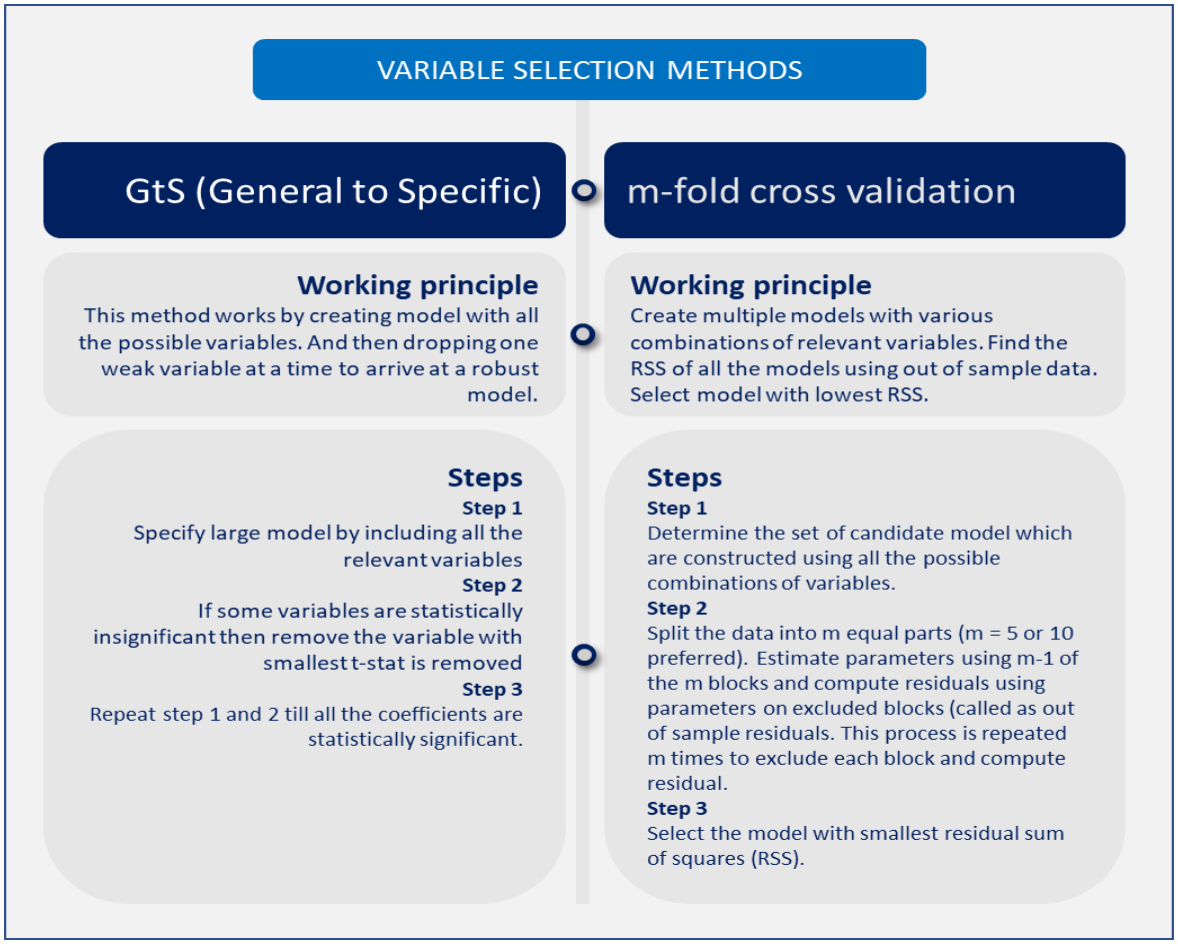
Effects of extraneous variables

- Does not affect the accuracy of the model and is not very serious problem but
- It increases variables in the model which results into decline in adj R2.
- When true coefficient is zero for added variable, it increases standard error in model. This creates the problem in financial data, where variables are high in correlation.

9.2.b Bias Variance trade off

Before the regression analysis begins, analyst must make a choice of including or excluding variables in the model. The inclusion of irrelevant variable increases variance and omitting relevant variable creates the bias in the model. The choice of inclusion or exclusion of a variable is trade-off between bias and variance. Bias variance trade-off is the fundamental challenge in variable selection.

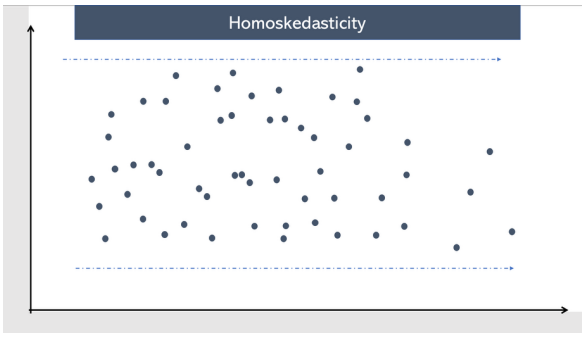
- Large models (with more explanatory variables) - Lower bias but higher variance
- Small models (with less explanatory variables) – Higher bias but lower variance

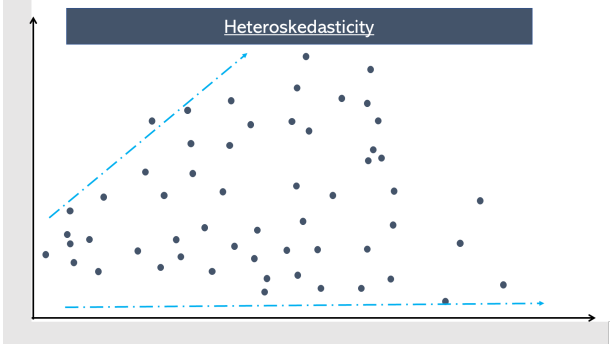


9.3 Heteroskedasticity

Homoskedasticity: One of the assumptions of OLS estimator is that the variance of error term e is constant conditional upon mean this is called homoskedasticity. When the residual is homoscedastic it means independent variable provides no information about the location of error term.

Heteroskedasticity. When the variance of error term is not constant across the observation, residuals are called heteroskedastic. In financial data it is very common for residual to be heteroskedastic.





Effects of Heteroskedasticity:

- Distribution of estimated parameters take different form.
- Consistency and unbiasedness of the OLS parameters is unaffected.

Detection using graphical method: Heteroskedasticity/ homoskedasticity can be detected using the residual plots.

Problem: Because heteroskedasticity does not affect the parameters hence fixes for parameters are not required. However, standard error is not reliable which affects the hypothesis testing. We cannot use standard error in hypothesis testing when heteroskedasticity is present in residuals.

Solution: The white standard error in place of standard error is used in hypothesis testing using t stat. However, for joint hypothesis testing F test is not easy to adjust for heteroskedasticity.

Approaches to modelling heteroskedastic data:

- Approach 1: Ignore heteroskedasticity when estimating the parameters and then use the heteroskedastic – robust (white) covariance estimators in hypothesis testing. This method often produces substantially less precise model parameters estimate when compared to method that directly address heteroskedasticity.
- Approach 2: Transforming data before modelling.
- Approach 3: Use weighted least squares (WLS) is a generalization of OLS.

(Note: For exam purpose remember the approaches used. Detailed approach is less likely to get testing in exam)

9.4 Multicollinearity

Multicollinearity occurs when one or more independent variables are highly correlated with others. Suppose a model has two explanatory variables, multicollinearity exists if one independent variable can be predicted with high R2 by another independent variable. Multicollinearity is not the same as perfect collinearity. Perfect collinearity is where two variables have a correlation of 1. Multicollinearity is a common issue in finance because data is influenced by the same market events. Multicollinearity does not violate any assumption so it does not cause technical difficulty in regression modelling (parameter estimation or hypothesis testing). But it is a problem in data modelling. When data are multicollinear, coefficients are jointly statistically significant but have low individual t statistics. This happens because the joint statistical analysis can detect some effect from the regressors as a group but cannot attribute the effect to a single variable.

Identifying multicollinearity:

The standard method to determine whether variable are excessively multicollinear can be detected using VIF (variance inflation factor). This measure compares the variance of regression coefficient on an explanatory variable X in two models 1) including only X and 2) including all explanatory variables.

VIF = $\frac{1}{1-Rj^2}$, where R^2 comes from regression of X on the other model. Value above 10 is considered excessive.

Variables with exceedingly high VIF should be excluded from the model.

Solution to multicollinearity:

- Ignore the multicollinearity because it is not technical problem in regression modelling.
- Identify multicollinear variable and to consider removing such from the model. Removing variable which is source of multicollinearity is difficult to identify.

9.5 Residual plots visualization

Residual plot is used to detect deficiencies in the model specification. An ideal model would have residuals that are not conditionally related to any of the explanatory variable. Residual should also be small in magnitude (± 4 s, where S2 is the estimated variance of the shock in the model). On residual plot estimated e is on Y axis and explanatory variable on X axis. Both outlier and model specification problem can be identified with these plots.

9.6 Outliers

Outliers are large values which affects the estimated coefficients largely on inclusion or exclusion from the data. Cook's distance measures the sensitivity of fitted values in regression to dropping a single observation.

$$D = \frac{\sum(\hat{Y}j - \hat{Y})^2}{KS^2}$$

Large value of Cook measure D> 1 indicates that observation j has rage impact on the estimated models parameters.

9.7 Which OLS is the best linear unbiased ESTIMATORS?

OLS is a linear estimator because both intercept and slope are linear function of Y. Under the assumption introduced, OLS estimators are BLUE (Best unbiased estimators). OLS achieves the smallest variance among any estimator that is linear and unbiased. OLS is the best estimator in the sense that any other LUE must have large variance.

Reading 10 Stationary Time Series

10 Stationary Time Series

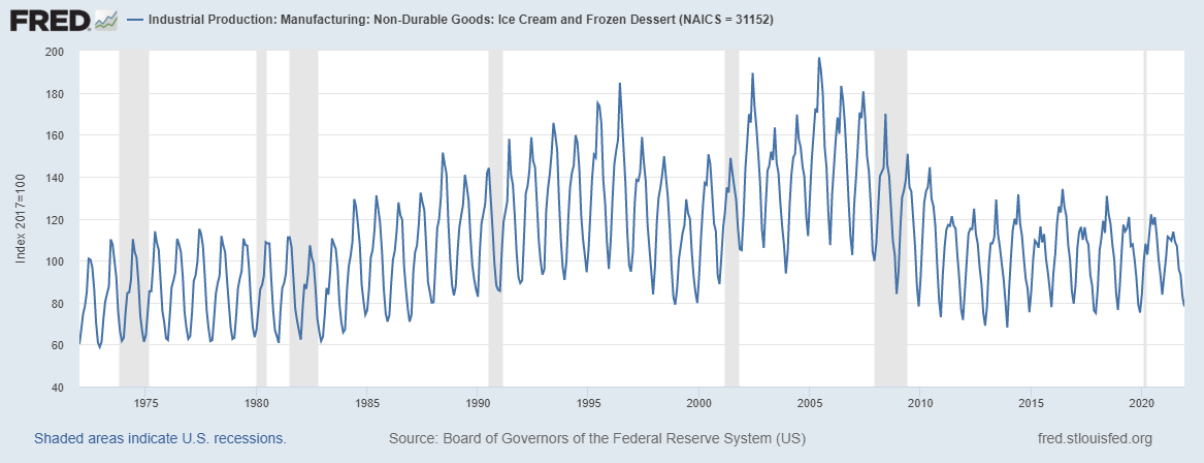
SCOPE OF THIS READING

This chapter develops covariance stationary time series models used in financial forecasting. It defines the conditions for covariance stationarity and introduces the autocovariance and autocorrelation functions, along with white noise processes. The chapter explains the structure and properties of AR, MA, and ARMA models, including the role of the lag operator and mean reversion, and derives the mean-reverting level. It discusses model identification using sample autocorrelation and partial autocorrelation functions, and diagnostic testing using the Box–Pierce and Ljung–Box Q-statistics. Finally, it explains how forecasts are generated from ARMA models, the importance of mean reversion for long-horizon forecasts, and how seasonality is incorporated within a stationary ARMA framework.

10.1 Time Series Introduction

In the earlier reading about regression analysis, we examined how the variable are related. We used cross sectional data for the regression model, which shows how one variable affects another variable. For example: How much weight is lost from doing intense cardio exercise and burning calories.

Time series on the other hand is the collection of observation drawn from different point in time. Consider example monthly ice cream sale of Blue Berry ice-cream. In this sales as a variable is analysed by keeping time on x axis. Following graph shows ice create and frozen deserts production data from 1972 to 2021.



Time series analysis aims to find relationship between the dependent variable (ice cream sale data or production data) and time (Time replaces the independent variable). Time series analysis is a simple process on the surface level but very mathematical at its core. For example, take the ice cream sales data. We can easily guess that the sales of ice cream go up in summers and go down in other months. But measuring the effect of this change is a mathematical process. Nowadays, we can use programming language and software to avoid the core mathematics. These tools will take care of all the mathematical part of time series analysis. We just need to know how to use these tools, which time series model to choose by looking at the type of time series, interpret the results from the model (given by the software) and make time series predictions using the software.

Note: This and the next reading will help us understand the different kinds of time series, the models that can analyze them, and how to choose the best model for our time series analysis. We will use some maths to learn the basics of the models for time series analysis, but don't worry if you find the maths challenging. GARP knows that, in practice, software handles the maths part and users' main role is to interpret the results. Maybe that's why, the learning objectives related to these readings requires interpretation skills and not the core mathematical skills.

Components of time series: Time series has 3 components called trend, seasonality (or cyclicality), and random error. Let's take a look at each component.

- **Trend**: Trend is the normal tendency of the observations in increasing or decreasing with time period. Time series may show periodic shift in trend just like we saw in previous graph (ice cream production is increasing for some time and then decreasing afterwards). Trend can also be linear or nonlinear. If the increase or decrease in time series is linear

function of time then it is linear trend. If trend shows curvature in movement, then it is nonlinear trend.

- **Seasonality**: Seasonal variation in time series is seasonality. In ice creme production graph, spikes are seasonal component i.e. increase is because of season of ice crème. Component can be called as seasonal if it shows similar variation for specific time in a year. If variation is not observed every year (say in the month of May every year) in same period, then that component cannot be called as seasonal. Similar to seasonal component, cyclical variation is the variation in regular interval of time but not observed every year. Example, GDP of a country falls in every 10th year.
- **Random white noise:** Random variation in time series is random noise. When random noise fulfils certain conditions, we call it a random white noise.

We will discuss all of these components in detail in this and next reading.

Additive vs Multiplicative Time Series

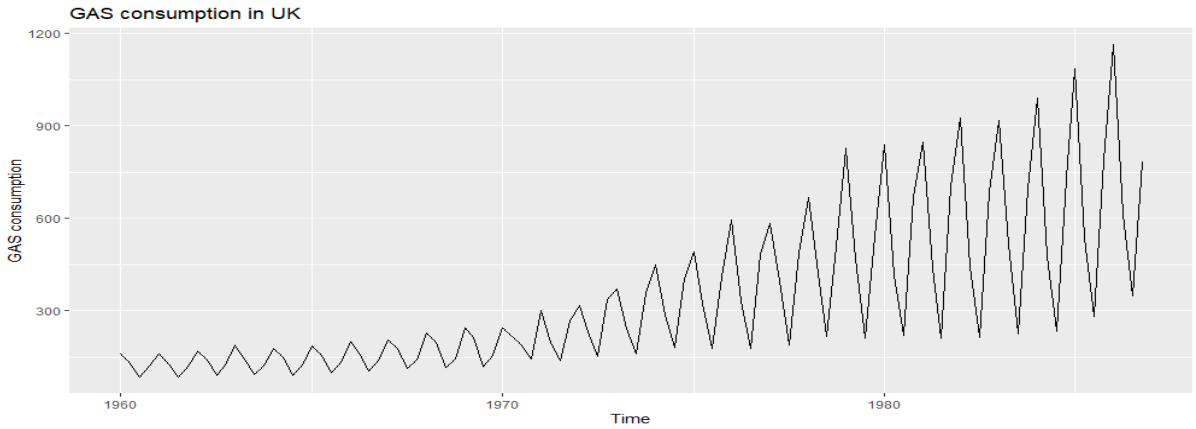
An additive time series model assumes that the components of the time series are independent and can be added together to obtain the observed values. For example, if x(t) is the monthly sales of a product, then x(t) = trend(t) + seasonality(t) + random noise(t), where trend(t) is the long-term increase or decrease in sales, seasonality(t) is the periodic variation due to seasonal factors, and random noise(t) is the unpredictable fluctuation that cannot be explained by the other components. An additive model is appropriate when the magnitude of the seasonal variation does not depend on the level of the trend.

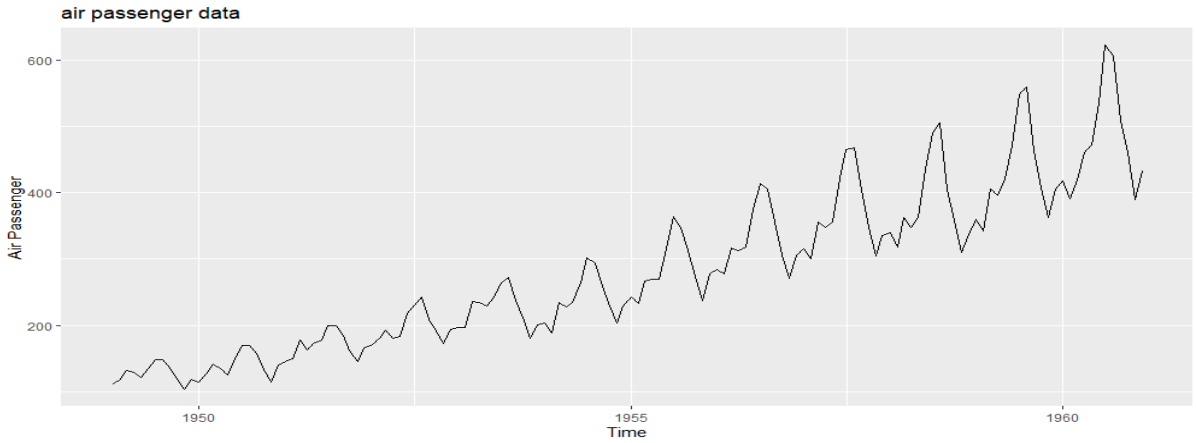
A multiplicative time series model assumes that the components of the time series interact with each other and can be multiplied together to obtain the observed values. For example, if x(t) is the monthly sales of a product, then x(t) = trend(t) * seasonality(t) * random noise(t), where the components are defined as before. A multiplicative model is appropriate when the magnitude of the seasonal variation increases or decreases proportionally with the level of the trend.

Simple additive time series model can be decomposed into

Observation x(t)= trend (t) + Seasonality(t) + Random white noise (t)

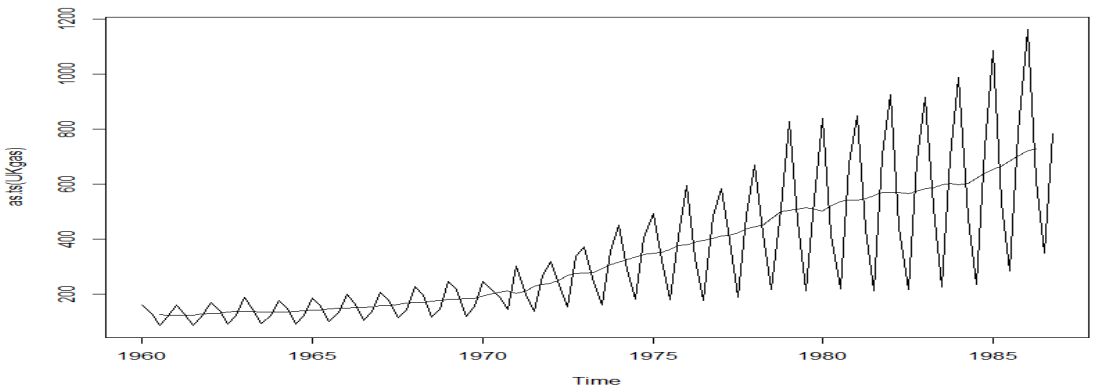
Following are some more examples of time series data in graphical format.



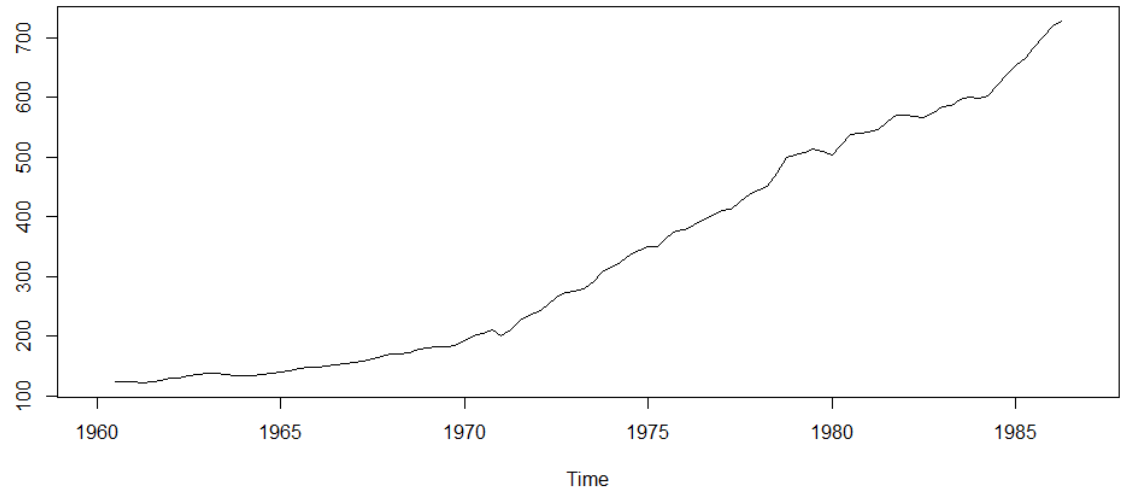


Let's take the example of GAS consumption in UK. This time series has all three components trend, seasonality and random error. Following are the graphs which shows each component separated from the time series. These components can be observed in original graph as well. We can see overall increasing trend in gas consumption and spikes are indicative of seasonality. However, random error cannot be directly observed in graph and needs separate consideration

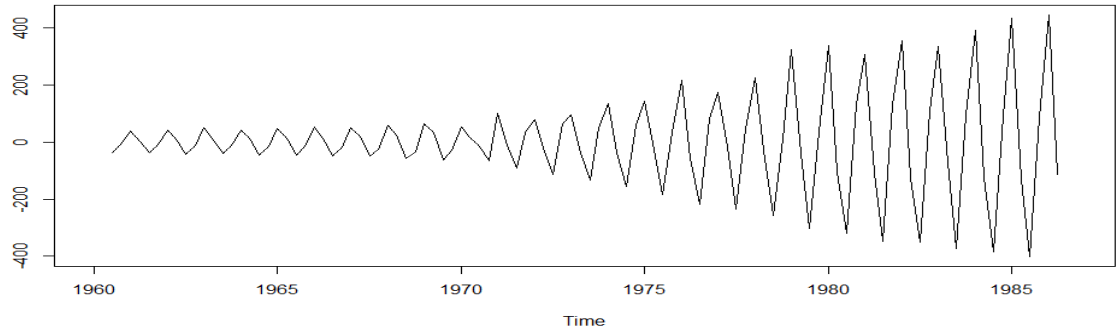
Overall trend (with original observations) in time series (UK Gas consumption)



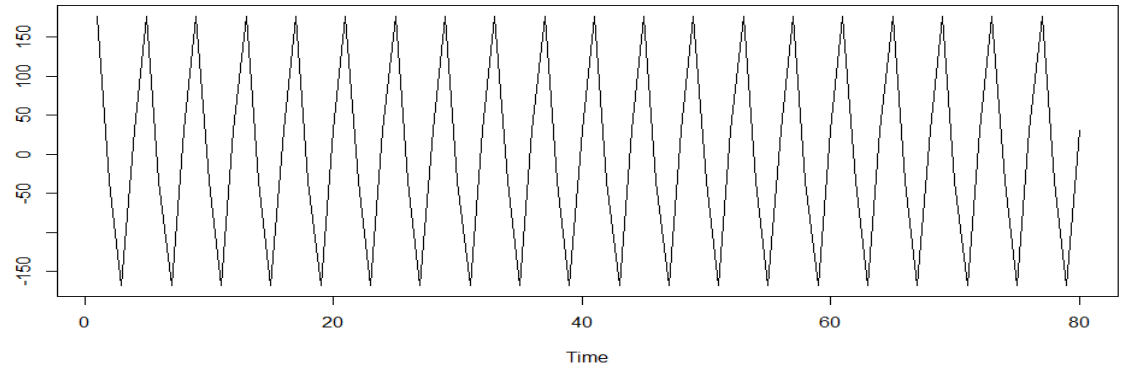
Only trend separated from time series



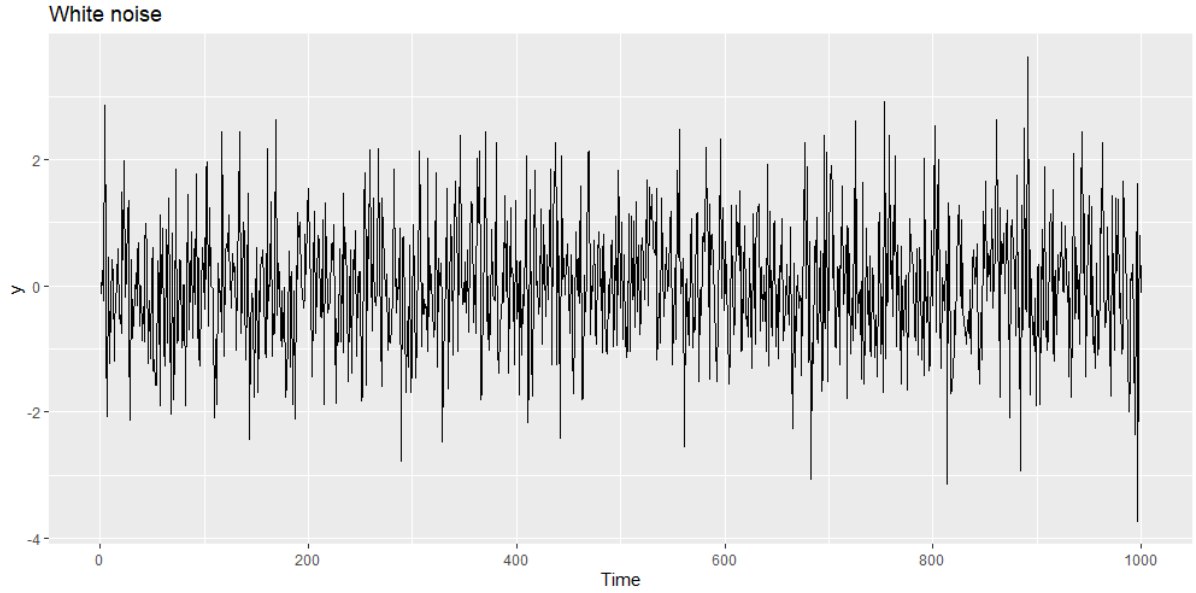
Time series after removing trend i.e. detrended time series



Only seasonal movement from the time series



White noise from time series (simulated and not from UKgas data)



Forecasting using time series: When the time series model is constructed using past observations, we can use it for forecasting of data using. For forecasting, there must be some form of stable relationship between the past data and future data. This relationship is called as stationarity of time series. In the following topics we will study about the meaning of stationarity

of time series and how to model stationary time series. In the next reading we will see how to model time series when stationarity property is violated.

10.2 Covariance stationary

Covariance stationary is the essential property of time series analysis using which time series models can be used in forecasting because it provides relationship of past data with future data. Time series is considered covariance stationary if it fulfills following conditions –

- Mean is constant and does not change over time
- The variance is finite and does not change over time
- The autocovariance is finite and, does not change over time and only depends on the distance between observations.

When the time series is covariance stationary, it means time series has constant relationship across time. Parameters estimated using nonstationary time series are more difficult to interpret and is also subjected to spurious relationship where relationship between observations will be statistically significant even if it has no relationship in reality. The covariance in the time series is the covariance of observations with its past observations which is called as autocovariance. Similar to correlation provides covariance, autocovariance provides the autocorrelation which is just standardized form of autocovariance.

10.3 Stochastic Process

Sequence of random variables is called stochastic process and denoted by $\{Y_t\}$, This reflects the fact that the sequence of random variable that are ordered in time. When forecasting with time series, ordering (sequence) is very important because it is based on past observations. The first order AR process is the example of stochastic process which we will learn soon.

First order AR model : $Y_t = \delta + \phi Y_{t-1} + e_t$

This model is similar to linear regression model which we discussed in previous readings, where δ is constant, ϕ is model parameter measuring strength of the previous observation at time t-1 with observation at time t.

In this chapter we will discuss linear stochastic processes. The process is linear in$\{e_t\}$ with mean zero stochastic process referred to as the shock. The intercept process is deterministic and the coefficients on the shocks are constant. In this chapter we will only cover models with constant deterministic factor δ (does not change with time). In the next reading we will discuss models which does not assume constant deterministic component and δ can change with time to accommodate the impact of trend and seasonal effects.

We will focus on linear process because linear process can be directly linked to linear models. We can use linear process for nonlinear processes because nonlinear processes have linear representation.

10.4 White noise

White noise is essential for time series.

White noise process: $\epsilon_t \sim WN\ (0, \sigma^2)$

Which indicates white noise is distributed with mean zero and variance. The moments, mean and variance in white noise process are not time dependent and hence process is covariance stationary. Shocks 'e' from the white noise process are used in data simulation.

White noise properties:

- **Mean zero.** This property offers convince of accommodating errors even if its mean is not zero. Nonzero mean errors can be translated into mean zero errors by subtracting mean value from all the values of error.
- **Constant and finite variance:** This assumption provides the support for next assumption
- **No Autocorrelation or autocovariance:** This assumption forces all autocorrelation in time series to be driven by model parameters and not shocks.

It is critical to test, whether the shocks from estimated model parameters are consistent with above properties.

Gaussian white noise process: If the random variables in white noise process are iid (independent and identically distributed) and normally distributed, then white noise process is called Gaussian White noise process.

Please note (imp for exam) white noise process itself does not assume any specific distribution. It is not at all necessary that nose is normally distributed for time series analysis. Normal distribution is assumption of Gaussian white noise process which is convenient to assume but not followed by financial assets.

Dependent white noise relaxes the iid assumption while maintaining the three key properties of white noise process. Dependent white noise can change with time. Example, volatility of financial asset moves in the regime of high volatility or low volatility. Dependent white noise can be different in these regimes while maintaining the three key properties of white noise process.

10.4.a Wolds Theorem

Wolds theorem provides justification for using linear process to model covariance stationary time series. It also establishes the role of white noise in covariance stationary process.

If Y_t is a mean zero covariance stationary process, then

$Y_t = \epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-1} + \ldots.$, where ψ terms are constants.

Wolds theorem states that this representation of a covariance stationary process is unique.

10.5 The Lag operator

Lag operator L shifts the time index of an observation, so that $LY_t = Y_{t-1}$. Following are the key properties of lag operators,

- Lag operator shifts the time index back one observation.
- Lag operator applied to constant is also constant
- $L^p = Y_{t-p}$
- Lag polynomials can be multiplied
- If the coefficient sin the lag polynomial satisfy some technical conditions, the polynomials can be inverted.

The concept of invertibility is useful in two cases

- AR process is only covariance stationary if its lag polynomial is invertible.
- Invertibility plays key role when selecting a unique model for a time series using Box Jenkins methodology.

10.6 Autocovariance and Autocorrelation

In time series value of observation y in period is correlated with its past values. The correlation with its lagged values (previous values) is called autocorrelation. As we know correlation is outcome of covariance, similarly autocorrelation is the function of autocovariance. When Yt is covariance stationary, the autocorrelation is defined as the ratio

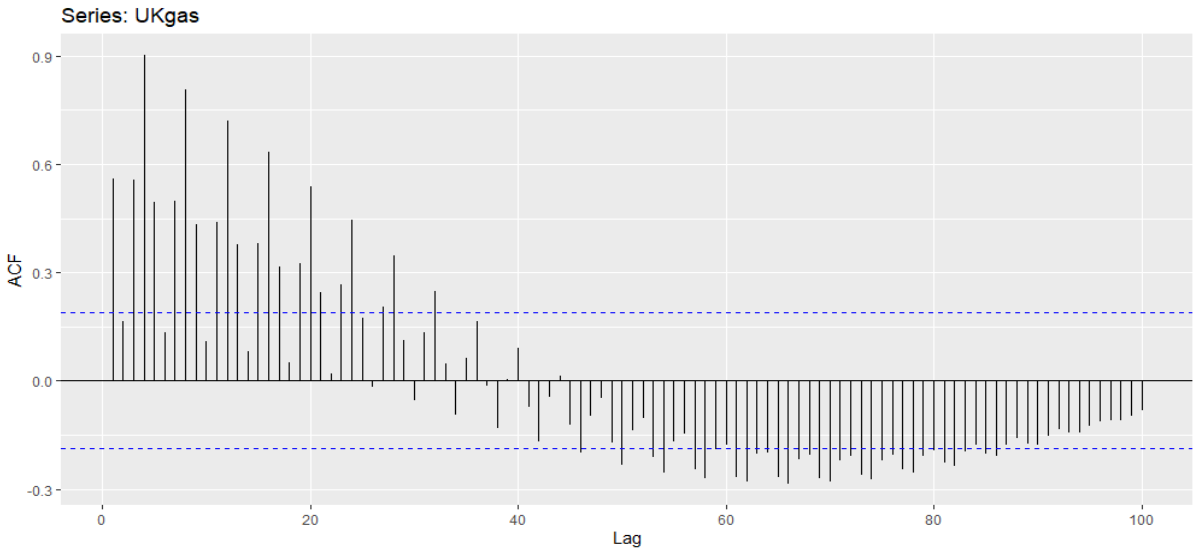
$$\rho_h = \frac{Cov\ (Y_t, Y_{t-h}}{\sqrt{var(Y_t)var(Y_{t-h})}}$$

Where h indicates the lagged period, for h = 3 means 3 periods back.

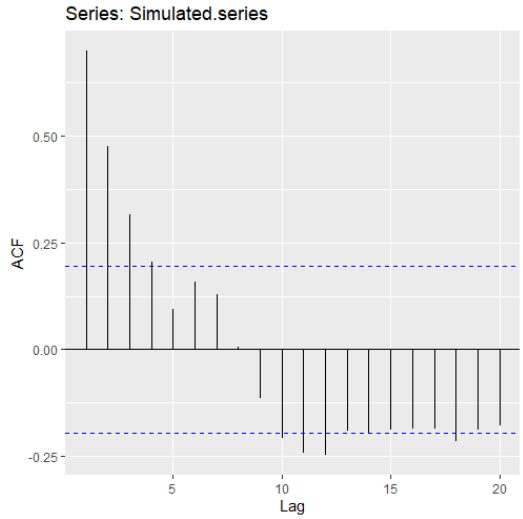
Similar to correlation, autocorrelation ranges from -1 to +1. Please note, autocorrelation is the function of h and not t because it does not depend upon time (stationary across time) and hence it is only well defined when the time series is stationary.

10.6.a Autocorrelation function (ACF) and Partial autocorrelation (PACF)

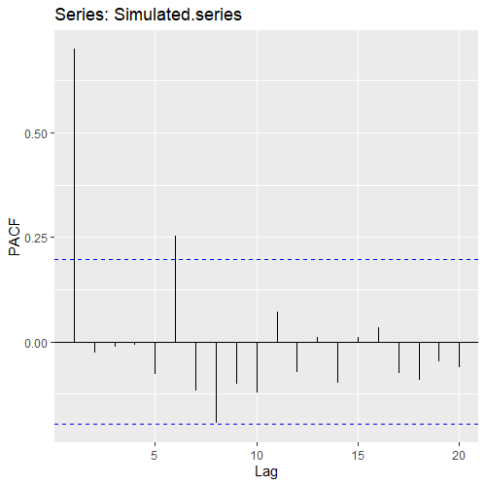
Autocorrelation function is similarly defined using the autocorrelation. ACF is the simples test of stationarity. In simple terms, ACF is series of autocorrelation of y with its lagged values. Following is the ACF plot, lags on x axis and ACF on y axis. ACF starts with taking autocorrelation of y with past period $y_{t-1,}$ for 2 lag autocorrelation, yt and y_{t-2} is used. Same process is opted for all the other lags which gives ACF plot. ACF decays to zero as h increase which can be seen from ACF plot. Constant decline in ACF is due to trend, and spikes in ACF are the result of seasonality.



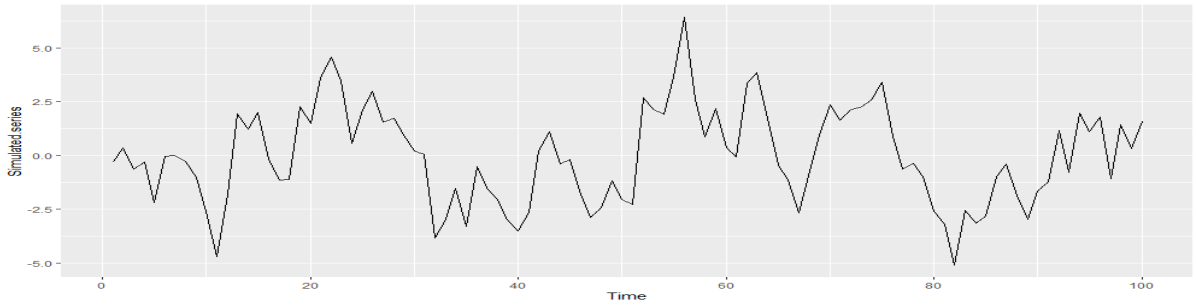
Partial autocorrelation PACF is similar to ACF except that each correlation controls for any correlation between observations of a earlier lags. PACF is nonlinear transformation of ACF and mainly used in model selection. ACF and PACF is used to determine the order of AR and MA model. In the firs lag, ACF and PACF measures the same correlation but PACF drops suddenly after first lag because the portion of correlation captured by first lag.



ACF plot of simulated series



PACF plot of simulated series



Simulated series

10.7 Autoregressive (AR) models

Autoregressive model takes the support of recent values of the stochastic process to its previous value. In simple terms it is regression of a variable Y_t with its lagged value Y_{t-1}. AR(1) is first order process,

$$Y_t = \delta + \phi Y_{t-1} + e_t$$

Where δ is intercept and ϕ is slope coefficient or parameter of AR and e is white noise shock. AR parameter determines the persistence of Yt. AR (1) is covariance stationary when $|\phi| < 1$ and non stationary when $|\phi| = 1$. When Yt is covariance stationary, the mean, variance and autocovariance are all constant.

10.7.a AR(p) process

The pth order AR process includes p lags of Y in the model. In simple term, it is regression of Y as dependent variable and Y_{t-1}, Y_{t-2} etc as independent variables.

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots\ldots \phi_p Y_{t-p} + e_t$$

AR process tends to move close to the mean. The sum of coefficients ϕ should be less than 1 is the necessary condition for stationarity. This is simple to check, simply take the sum of all coefficients. If sum is more than 1 then process is not stationary.

ACF / PACF of an AR(p) process and AR(1) shares similar structure. Example: ACF of an AR(p) process decays as the length of lag increases and may oscillate. PACF of AR(q) process shows sharp drop at p lags. Hence, PACF of AR(1) process cuts off after just one lag.

10.7.b Yule-Walker equation

The mathematics that governs the AR model is the Yule Walker equation. Yule Walker equation connects the parameters of AR model to the covariance function of the process. Hence, model parameters can be estimated from the covariance of the time series. This equation provide an expression that relates the parameters of an AR to the autocovariance of AR process.

Note: Yule Walker equation provides the derivation behind the AR model which is complicated. From exam perspective, it is highly unlikely to get tested on this equation. If you are interested in knowing more about the Yule Walker equation, please read GARP book page number 167.

10.8 Moving Average (MA) Model

All the variation in time series is driven by shocks of various types, suggests the possibility of modeling time series directly as a distributed lags of current and past shocks is the moving average process. MA(1) process is first order moving average process denoted by

MA(1) process: $Y_t = \mu + \theta \epsilon_{t-1} + \epsilon_t$

Where error term is white noise process. The Y_t depends on both the contemporaneous shock ϵ_t and previous shock ϵ_{t-1}. The parameter θ is weight and determines the strength of the effect of the previous shock. The μ is the mean of the process. This model equation has two implications

- When θ is positive MA(1) is persistent because two values are positively correlated.
- When θ is negative MA(1) is mean reverting because effect of previous shock is reverted.

Moving averages are always covariance stationary. MA(1) has limited memory, because only shocks of previous period impacts the current value. Any MA(1) has exactly one non zero correlation and ACF is zero for h≥ 2(i.e. sharp cutoff of autocorrelation function). The PACF of MA(1), is more complex and has non zero values at all lags. This is inverse of what AR(1) produce.

MA(q)

The general finite order moving average process of order q is generalization of MA(1).

$$Y_t = \mu + \theta_1 \epsilon_{t-1} + \ldots\ldots + \theta_q \epsilon_{t-q}$$

Where, all the shocks are white noise and have zero mean.

ACF is always zero for lags larger than q and the PACF is non zero for all lags.

Important differentiation of ACF and PACF in MA and AR processes.

	PACF	ACF
AR model	Cuts off sharply	Oscillates and decays slowly towards zero.
MA model	Decays slowly	Cuts off sharply

10.9 Autoregressive Moving Average (ARMA) models

ARMA is combined model in order to obtain better and parsimonious approximation. ARMA(1,1) indicates AR of first order and MA of first order.

ARMA(1,1) model: $Y_t = \delta + \phi Y_{t-1} + \theta \epsilon_{t-1} + \epsilon_t$

Autocovariance function is complex in ARMA(1,1). ACF decays as h increases and oscillates if ϕ<0. This is consistent with ACF of AR process. PACF decays slowly towards zero which is consistent with MA process. The slow decay of both ACF and PACF is key feature of ARMa model.

ARMA(1,1) is covariance stationary if $|\theta|$ <1. The MA coefficients is not involved in determining the stationarity of the this model.

ARMA(p,q) process is combination of AR(p) and MA(q) process. ARMA(p,q) is also covariance stationary, if AR component stationary. ACFs of ARMA process are more complicated than pure AR and MA models. ACF and PACF decays slowly for ARMA(p,q) as well.

10.10 Sample Autocorrelation

Sample autocorrelations and partial autocorrelations are used to build and validate ARMA models. These tools are first applied to the data to understand the dependence structure and to select a set of candidate models (by decided order of model). Then these tools are applied to estimate residuals to decide whether they are consistent with the key assumption of errors are white noise.

10.10.a Join Test of Autocorrelation

The autocorrelation in the residuals from the ARMA model can be evaluated graphically or from the formal tests (model/formula based). For graphical examination of a fitted model includes plotting residuals or ACF and PACF of residuals. Analyzing residuals of fitted model using graphical methods is sometimes challenging, hence, formal testing can be used with graphical methods.

Two tests which are used for joint testing of autocorrelations for validating a model. They both test the joint null hypothesis that all of the autocorrelations are simultaneously zero.

H_0: $\rho_1 = \rho_2 = \rho_3 = \ldots. = \rho_h = 0$

H_a: $\rho_j \neq 0$ *for at least one is non zero.*

Values of the test statistics larger than the critical value indicate that the autocorrelations are not zero.

The Box-Pierce Test (when sample size is large)

The Box-Pierce test statistics is the sum of the squared autocorrelations scaled by the sample size T.

$$Q_{BP} = T \sum_{i=1}^{h} \hat{\rho}^2$$

Q_{BP} = Chi squared statistics (h degrees of freedom)

Ljung Box statistics (When sample size is small)

Ljung Box statistics is version of Box Pierce statistics that works better in smaller sample size. When sample size is modest, the finite sample distribution of the Ljung Box when the null is true is close to the asymptotic chi squared distribution. Therefore it is preferred method to test multiple autocorrelations.

10.11 Model building and selection

Initial model building of AR, MA or ARMA requires review of ACF and PACF. The main consideration in model building is choice of total lags p for AR and q for MA. First, ACF and PACF is analyzed. The slow decay in ACF indicates that the model is good fit for AR and slow decay in PACF indicates that the model is good fit for MA component. Using these steps, suitable candidate model is selected with specific lags.

Once initial set of model is identified, we need to check the measure of fit for these models. Measure of fit is Mean squared error (MSE) of the model. Smaller value means model is better fit. Problem with this residual analysis is that, adding more lags will always lower Mean squared error. Hence only aiming for minimizing mean squared error is not ideal solution. This situation is similar to regression models we discussed previously i.e. adding extra variable increases R^2 which increases complexity in model. The solution is similar here, penalizing MSE for added lags. These measures are information criteria (IC) - Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). Both the IC balances the bias variance tradeoff. Penalty cost in AIC is constant 2, but for BIC it is variable and increases gradually with time. There are two implications to this –

- The BIC always selects a model that is no longer in lag than the model selected by the AIC.
- The BIC is a consistent model selection criterion. i.e. true model is selected as T increases.
- The AIC behaves like a model selection methodology which can lead to selection models that are too large.
- The BIC is similar but variables that are not needed are always excluded.

10.12 Box Jenkins

Two models can be different in parameters but equal in ACF and PACF. The Box Jenkins methodology provides two principles to select among the equivalent models.

- Parsimony: Always choose model with lesser number of parameters
- Invertibility: When choosing parameters in MA process (also include ARMA), always select parameter values so that the MA coefficients are invertible.

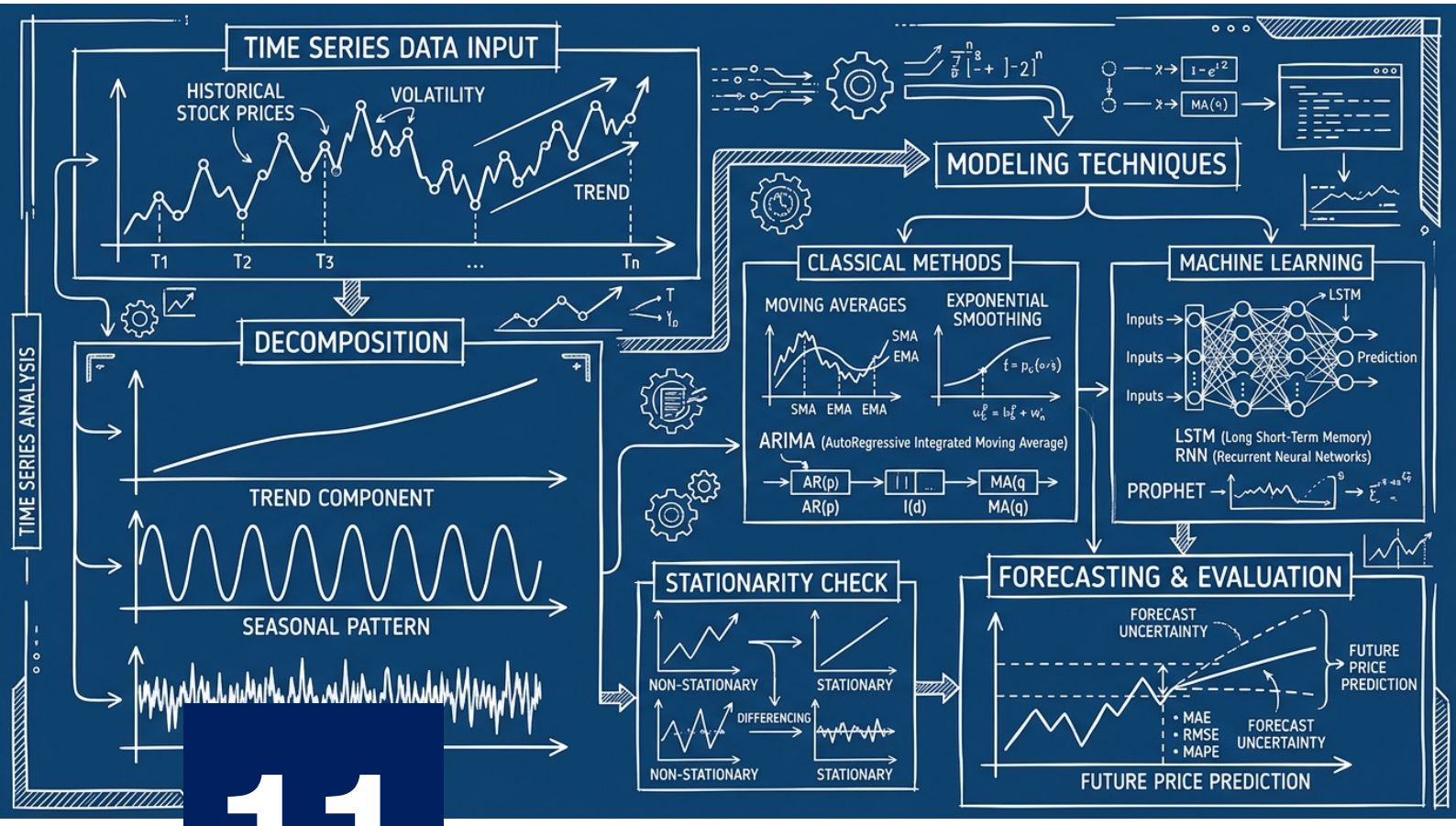
10.13 Seasonality

Seasonality is the product of human behavior, like people eat more ice cream in summer, people travel more in the month of December but must occur on annual basis. Seasonality can be a constant(deterministic) or changing (stochastic). Series with deterministic seasonality are non-stationary. The seasonal component uses lags at the seasonal frequency, while the short term component uses lags at the observation frequency. A seasonal ARMA combines these two components. In practice seasonal components are usually restricted to one lag because the

precision of the parameters related to the seasonal components depends on the number of full seasonal cycles in the sample.

Model selection in seasonal time series is identical to the selection in non-seasonal time series. Seasonal AR have slow decaying ACF and a sharp cutoff in the PACF. Seasonal MAs have opposite pattern, where the PACF slowly decays and the ACF drops off sharply.

Reading 11 Non-Stationary Time Series



11 Non-Stationary Time Series

SCOPE OF THIS READING

This chapter develops regression-based approaches for modeling time series behavior. It distinguishes between linear and nonlinear time trends and explains how seasonality can be incorporated using regression techniques. The chapter introduces the concept of a unit root and the random walk hypothesis, discusses the challenges posed by non-stationary time series, and explains how to test for the presence of a unit root. It also develops multi-step ahead forecasting for seasonal time series and shows how to estimate trend values and construct interval forecasts.

Any time series which is not covariance stationary is known as non-stationary time series. The series that we want to forecast vary over time, and we often attribute that variation to unobserved underlying components, such as trends, seasonal and unit roots. Modeling and forecasting nonstationary time series can be very challenging, hence we need to opt some procedures which splits these components from the time series and then reintroduced in the model for forecasting. In this reading we will discuss each of these components in detail and how model time series with these components.

11.1 Trends in time series

Trend is gradual, lasting, change in the variables that we want to model and forecast. In finance and economics trend is mostly seen in time series. We'll concentrate here on models of deterministic trend, in which the trend changes in a completely certain way. Trends can be linear or non linear.

Linear trend means it increases or decreases like a straight line. Example of linear time trend is

$$Y_t = \delta_0 + \delta_1 x\ time + \epsilon_t$$

Slope δ_1 in the model is linear function and hence this is linear trend. This same slope makes trend series nonstationary because mean is time dependent. If slope is > 0 then trend is increasing with time and slope <0 means trend is decreasing in time.

Nonlinear trend or curved trend is when the increase or decrease in trend is at increasing or decreasing rate. It is not necessary for trends to be linear. Quadratic trend models can capture nonlinearities.

$$Y_t = \delta_0 + \delta_1 x\ time + \delta_2 x\ time^2\ \epsilon_t$$

Linear trend is the special case of nonlinear trend where δ_2 is equal to zero. Higher order or polynomials are sometimes entered but for smooth trend it is better to use lower order polynomials.

Both the models used above uses growth factor in trend. In finance growth factor is sometimes not appropriate. Assume the time series of stocks with negative growth factor, which leads to negative values. Hence it is better to use growth rate instead of growth factor. Growth rate can be introduced in the model with the help of logarithms. Trend which appears nonlinear in levels but linear in logarithms, is called exponential trend, or log linear trend, and is very common in finance and economics. That's because economic variables often display roughly constant growth

Rates.

$$Ln\ (Y_t) = \delta_0 + \delta_1 x\ time + \delta_2 x\ time^2\ \epsilon_t$$

R^2 in trending series is always high and inevitable and is not suitable measure for trend time series. Instead of R^2, residual diagnostic or other formal tests are used to assess model strength.

11.2 Seasonality

If a time series is observed at monthly or quarterly intervals (or even weekly or daily), it may exhibit seasonality. For example, monthly housing starts in the Midwest are strongly influenced

by weather. Although weather patterns are somewhat random, we can be sure that the weather during January will usually be more inclement than in June, and so housing starts are generally higher in June than in January. One way to model this phenomenon is to allow the expected value of the series, y_t, to be different in each month. As another example, retail sales in the fourth quarter are typically higher than in the previous three quarters because of the Christmas holiday. Again, this can be captured by allowing the average retail sales to differ over the course of a year. This is in addition to allowing for a trending mean. For example, retail sales in the most recent first quarter were higher than retail sales in the fourth quarter from 30 years ago, because retail sales have been steadily growing. Nevertheless, if we compare average sales

within a typical year, the seasonal holiday factor tends to make sales larger in the fourth quarter. Even though many monthly and quarterly data series display seasonal patterns, not all of them do. For example, there is no noticeable seasonal pattern in monthly interest or inflation rates. In addition, series that do display seasonal patterns are often seasonally adjusted before they are reported for public use. A seasonally adjusted series is one that, in principle, has had the seasonal factors removed from it.

Sometimes, we do work with seasonally unadjusted data, and it is useful to know that simple methods are available for dealing with seasonality in regression models. We can include asset of seasonal dummy variables to account for seasonality in the dependent variable, the independent variables, or both. The approach is simple. Suppose that we have monthly data, and we think that seasonal patterns within a year are constant across time. For example, since Christmas always comes at the same time of year, we can expect retail sales to be, on average, higher in months late in the year than in earlier months. Or, since weather patterns are broadly similar across years, housing starts in the Midwest will be higher on average during the summer months than the winter months.

A general model for monthly data that captures this phenomenon is

$$Y_t = \beta_0 + \delta_1 feb_t + \delta_2 march_t + \delta_3 apr_t + \ldots. + \epsilon$$

Where, Feb_t, $Marh_t$, … are dummy variables indicating whether time period corresponds to the appropriate month.

Now let's construct seasonal dummy variables, which indicate which season we're in. If, for example, there are four seasons, we create:

D1 = (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, ...)

D2 = (0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, ...)

D3 = (0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, ...)

D4 = (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, ...).

D1 indicates whether we're in the first quarter (it's 1 in the first quarter and zero otherwise), D2 indicates whether we're in the second quarter (it's 1 in the second quarter and zero otherwise), and so on. At any given time, we can be in only one of the four quarters, so one seasonal dummy is 1, and all others are zero.

The pure seasonal dummy model is

$$y_t = \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t.$$

We are just regressing on an intercept, but we allow for a different intercept in each season. Those different intercepts, the, are called the seasonal factors; they summarize the seasonal pattern over the year.

Instead of including a full set of s seasonal dummies, we can include any s-1 seasonal dummies and an intercept. Then the constant term is the intercept for the omitted season, and the coefficients on the seasonal dummies give the seasonal increase or decrease relative to the omitted season. In no case, however, should we include seasonal dummies and an intercept. Including an intercept is equivalent to including a variable in the regression whose value is always one but note that the full set of s seasonal dummies sums to a variable whose value is always one.

11.3 Forecasting with seasonality and trend (h-step-ahead forecast)

In many forecasting situations, however, more than one component is needed to capture the dynamics in a series to be forecast. Here we assemble our tools for forecasting trends, seasonals; we use regression on a trend and seasonal dummies, and we capture cyclical dynamics by allowing for ARMA effects in the regression disturbances.

Trend may be included as well, in which case the model

$$y_t = \beta_1 TIME_t + \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t.$$

The idea of seasonality may be extended to allow for more general calendar effects. "Standard" seasonality is just one type of calendar effect. Two additional important calendar effects are holiday variation and trading-day variation. Holiday variation refers to the fact that some holidays' dates change over time. That is, although they arrive at approximately the same time each year, the exact dates differ. Easter is a common example. Because the behavior of many series, such as sales, shipments, inventories, hours worked, and so on, depends in part on the timing of such holidays, we may want to keep track of them in our forecasting models. As with seasonality, holiday effects may be handled with dummy variables. In a monthly model, for example, in addition to a full set of seasonal dummies, we might include an "Easter dummy," which is 1 if the month contains Easter and 0 otherwise.

Trading-day variation refers to the fact that different months contain different numbers of trading days or business days, which is an important consideration when modeling and forecasting certain series. For example, in a monthly forecasting model of volume traded on the London Stock Exchange, in addition to a full set of seasonal dummies, we might include a trading day variable, whose value each month is the number of trading days that month.

Allowing for the possibility of holiday or trading day variation gives the complete model

$$y_t = \beta_1 TIME_t + \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t.$$

$$y_t = \beta_1 TIME_t + \sum_{i=1}^{s} \gamma_i D_{it} + \sum_{i=1}^{v_1} \delta_i^{HD} HDV_{it} + \sum_{i=1}^{v_2} \delta_i^{TD} TDV_{it} + \varepsilon_t,$$

where the HDVs are the relevant holiday variables (there are v1 of them) and the TDVs are the relevant trading day variables (here we've allowed for v2 of them, but in most applications v2=1 will be adequate). This is just a standard regression equation and can be estimated by ordinary least squares.

Once the model is constructed, we can expand this model for forecasting of h step ahead (T+h) time period. This is out of sample forecast because the forecast is for period which is not observed or collected in sample. The full model with h step ahead is

$$y_{T+h} = \beta_1 TIME_{T+h} + \sum_{i=1}^{s} \gamma_i D_{i,T+h} + \sum_{i=1}^{v_1} \delta_i^{HD} HDV_{i,T+h} + \sum_{i=1}^{v_2} \delta_i^{TD} TDV_{i,T+h} + \varepsilon_{T+h}.$$

11.4 Random Walk and unit roots

11.4.a Random walk

There are many financial time series in which the changes follow a random pattern. We discuss these "random walks" in this section. A random walk is one of the most widely studied time-series models for financial data. A random walk is a time series in which the value of the series in one period is the value of the series in the previous period plus an unpredictable random error. A random walk can be described by the following equation

$$x_t = x_{t-1} + \epsilon_t, E(\epsilon_t) = 0, E(\epsilon_t^2) = \sigma^2, E(\epsilon_t \epsilon_s) = 0 \text{ if } t \neq s$$

time series xt is in every period equal to its value in the previous period plus an error term, Xt , that has constant variance and is uncorrelated with the error term in previous periods.

When the time series is random walk, variance increases with time and hence this series is not covariance stationary. Hence it is not possible to model this series with standard regression models such as AR, MA or ARMA.

11.4.b Unit root

A random walk is a special case of what is known as a unit root process. The name comes from the fact that $r_1 = 1$ in the AR(1) model. A more general class of unit root processes is generated

If a series has a unit root, its autocorrelation function isn't well-defined in population, because its variance is infinite. But the sample autocorrelation function can of course be mechanically computed in the usual way, because the computer software doesn't know or care whether the data being fed into it have a unit root. The sample autocorrelation function will tend to damp extremely slowly; loosely speaking, we say that it fails to damp. The reason is that, because a random walk fails to revert to any population mean, any given sample path will tend to wander above and below its sample mean for long periods of time, leading to very large positive sample autocorrelations, even at long displacements. The sample partial autocorrelation function of a unit root process, in contrast, will damp quickly: it will tend to be very large and close to one at displacement 1, but will tend to be smaller and decay quickly thereafter.

11.4.c Challenges in modelling time series with unit roots

If the time series contains unit roots, then it cannot be directly modelled because –

- Unit root time series is not mean reverting.
- It shows spurious relationship among the different unit root series
- Correct ARMA model cannot be selected because estimated parameters follow Dicky fuller distribution (size dependent and time trending).

If a time series appears to have a unit root, how should we model it? One method that is often successful is to first difference the time series and try to model the first-differenced series as an autoregressive time series.

The first difference of a time series is the series of changes from one period to the next. If Y_t denotes the value of the time series Y at period t, then the first difference of Y at period t is equal to Y_t-Y_{t-1}.

11.5 Dicky Fuller Test

Augmented Dicky Fuller (ADF) test is used to detect if the time series is unit root or not. Dickey-Fuller test is a unit root test that tests the null hypothesis that α=1 in the following model equation. alpha is the coefficient of the first lag on Y. Null Hypothesis (H0): alpha=1. The theory used to obtain the asymptotic critical values is rather complicated and is covered in advanced texts on time series econometrics

An ADF test is implemented using an OLS regression where the difference of a series is regressed on its lagged level, relevant deterministic terms, and lagged differences. The general form of an ADF regression is

$$\Delta Y_t = \underbrace{\gamma Y_{t-1}}_{\text{Lagged Level}} + \underbrace{\delta_0 + \delta_1 t}_{\text{Deterministic}} + \underbrace{\lambda_1 \Delta Y_{t-1} + \ldots + \lambda_p \Delta Y_{t-p}}_{\text{Lagged Differences}}$$

The ADF test statistic is the t-statistic of y. To understand the ADF test, consider a implementing a test with a model that only includes the lagged level:

$$\Delta Y_t = \gamma Y_{t-1} + \epsilon_t$$

so that the value of y is 0 when the process is a random walk. Under the null H0: y = 0, Yt is a random walk. The alternative is H1 : y < 0, which corresponds to the case that Yt is covariance stationary. Note that the alternative is one-sided, and the null is not rejected if y > 0. Positive values of y correspond to an AR coefficient that is larger than 1, and so the process is explosive and not covariance stationary. Implementing an ADF test on a time series requires making two choices: which deterministic terms to include and the number of lags of the differenced data to use. The number of lags to include is simple to determine—it should be large enough to absorb any short-run dynamics in the difference Yt. The lagged differences in the ADF test are included to ensure that error term is white noise process. The recommended method to select the number of lagged differences is to choose the lag length to minimize AIC. The maximum lag length should be set to a reasonable value that depends on the length of the time series and the frequency of sampling.

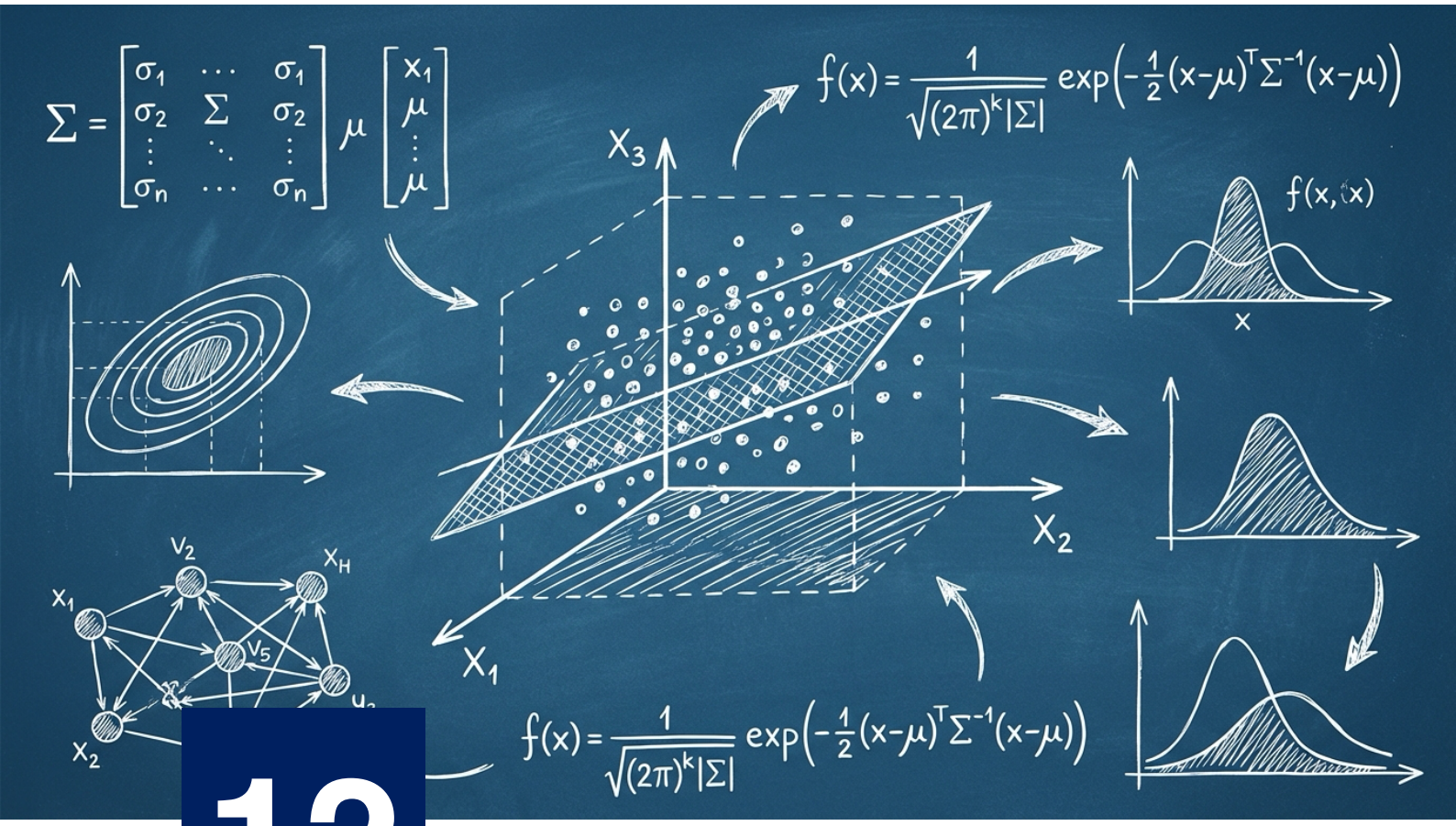
Recall that the AIC tends to select a larger model than criteria such as the BIC. This approach to selecting the lag length is preferred because it is essential that the residuals are approximately white noise, and so selecting too many lags is better than selecting too few. Ultimately, any reasonable lag length selection procedure—IC-based, graphical, or general-to-specific selection—should produce valid test statistics and the same conclusion.

The included deterministic terms have a more significant impact on the ADF test statistic. The DF distribution depends on the choice of deterministic terms. Including more terms skews the distribution to the left, and so the critical value becomes more negative as additional deterministic terms are included. For example, the 5% critical values in a time series with 250 observations are —1.94 when no deterministic terms are included, —2.87 when a constant is included, and —3.43 when a constant and trend are included. All things equal, adding additional deterministic terms makes rejecting the null more difficult when a time series does not contain a unit root. This reduction in the power of an ADF test suggests a conservative approach when deciding which deterministic trends to include in the test.

On the other hand, if the time series is trend-stationary, then the ADF test must include a constant. If the ADF regression is estimated without the constant, then the null is asymptotically never rejected, and the power of the test is zero. Avoiding this outcome requires including any relevant deterministic terms. The recommended method to determine the relevant deterministic terms is to use t-statistics to test their statistical significance using a size of 10%. Any deterministic regressor that is statistically significant at the 10% level should be included. If the trend is insignificant at the 10% level, then it can be dropped, and the ADF test can be rerun including only a constant. If the constant is also insignificant, then it too can be dropped, and the test rerun with no deterministic components. However, most applications to financial and macroeconomic time series require the constant to be included.

When the null of a unit root cannot be rejected, the series should be differenced. The best practice is to repeat the ADF test on the differenced data to ensure that it is stationary. If the difference is also non-stationary (i.e., the null cannot be rejected on the difference), then the series should be double differenced. If the double-differenced data are not stationary, then this is an indication that some other transformation may be required before testing stationarity. For example, if the series is always positive, it is possible that the natural log should be used instead of the unadjusted data.

Reading 12 Measuring Returns. Volatility and Correlation



12 Measuring Returns Volatility and Correlation

SCOPE OF THIS READING

This chapter develops statistical properties of financial returns. It explains the calculation and conversion of simple and continuously compounded returns, and distinguishes among volatility, variance rate, and implied volatility. The chapter evaluates the limitations of relying solely on mean and variance for non-normal distributions and introduces the Jarque–Bera test to assess normality. It discusses the power law in modeling heavy-tailed distributions and formalizes covariance and correlation, distinguishing linear correlation from broader dependence. Finally, it analyzes correlation properties under a one-factor model with normally distributed variables.

12.1 Introduction

Asset return volatility change from one period to another have important implication for risk management. As the volatility increases, probability of the loss on asset increases. In this reading we will learn why asset return distributions deviate from normality (i.e. not normal distribution). The return distribution is fat tailed is the outcome of time varying volatility. Returns can also be skewed which also makes it non normal distribution.

In the earlier readings we discussed correlation measure, which is very important for portfolio optimization, because optimization heavily depends on the correlation between the assets. In the earlier readings we used Pearson's correlation coefficient measure which is useful when assets show some form of linear correlation. However, two assets might not be linearly correlated if the Pearson's correlation coefficient measure is zero. This does not mean that assets are not correlated at all. There may be some form of correlation like nonlinear correlation. In this reading our main focus is on these measures of nonlinear correlation between two assets.

12.2 Returns

Simple return is the calculated using simple method and it gives us the effective return. Consider an investor purchased a security at time t-1 at price P_{t-1} and sold it at time t at P_t. To calculate simple return R_t

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Illustration: A trader purchased stock at $100 and sold it at $120 after 1 year. Return earned by a trader is

$R_t = \frac{120-100}{100} = 20\%$

We can use same example to calculate continuously compounded return. By using formula

$$\ln\left(\frac{P_t}{P_{t-1}}\right)$$

To calculate continuously compounded return (log returns),

$$R_t = Ln\left(\frac{120}{100}\right) = 0.18.232 = 18.232\%$$

Please note return in both the cases (effective and continuous) results into same earning for trader. The difference is due to quotation form. If we convert 20% effective annual return into continuously compounded return we will get the same result of 18.232%.

TI BA II calculator (Effective to continuous) : (1+0.20) > ln = 18.232%

Similarly we can also convert continuous into effective by 0.18232 >2nd > exp = 1.20 – 1 = 20%

We prefer log returns because to calculate multiple period returns we can simply add log returns. Log returns are approximation of effective return but log returns are less accurate for larger return values (eg: 30% return). In value terms log returns is always less than simple effective returns but this does not mean lower earning for investor as we discussed above. Also note that lower limit for simple return is 100%. This is because investor can maximum lose is portfolio which is 100% loss on portfolio or say return of – 100%.

12.3 Volatility and risk

As we discussed in the previous reading volatility is measured by standard deviation of the returns. We can use volatility to calculate return on a financial assets by using simple formula

$$R_t = \mu + \sigma e_t$$

Where e_t is shock with mean zero and variance of 1. The shock is assumed to be independent and identically distributed iid across observations and also normally distributed. This assumption also means that the returns are normally distributed. However, for most financial assets returns are not completely true.

Volatility is measured using standard deviation (σ). If the returns are computed using daily closing price i.e. if we use daily returns for calculation of standard deviation then the result of this calculation is daily volatility. There is also calculation period dependency in volatility. If we use 100 days return data then the result is daily volatility based on 100 days data which might differ if number of days taken for calculation are different.

12.3.a Time Scaling of volatility

Time scaling of volatility is especially important. We will use this concept in VaR section in Book 4. Assume a daily return volatility of 2%. To calculate annual volatility assuming 250 days, we need to simply multiply it by square root of 250.

$$Annual\ volatiliy = 0.02\ x\ \sqrt{250} = 31.62\%$$

One might ask why to take square root of time (days in this case). The simple answer is volatility σ is the square root term of variance σ^2. Because volatility is the square root term of variance, it should be scaled by root of time to match the square root of variance. Please note, no matter what the scenario is, volatility is always multiplied by root of time and not time directly.

Illustrations 1:

Assume annual volatility of 24%, then monthly volatility (12 months in a year).

$$Monthly\ volatiliy = \frac{0.24}{\sqrt{12}} = 31.62\%$$

In this case scaling is downwards i.e. from annual volatility to monthly hence we need to divide volatility by square root of time. When the scaling is upward we need to multiply volatility by square root of time.

12.3.b Implied volatility (Read this section after reading BSM reading from Book 4)

Implied volatility is an alternative measure of calculating volatility using BSM model. We know that, in BSM model call price, spot price, interest rate, strike price and time are observable factors (can be observed in market). The only variable which is not observable is volatility. We can use BSM model, to calculate volatility. Because this volatility is implied by BSM model. This implied volatility by structure is an annual value and so it does not need to be transformed further. The BSM option pricing model uses several simplifying assumptions that are not consistent with actual market. Also the model assumes the variance / volatility is constant over time.

The VIX Index is another measure of implied volatility that reflects the implied volatility on the S&P 500 over the next calendar 30 days constructed using options with a wide range of strike

price. The VIX method has been extended to many other assets, including other key equity indices, stocks, crude oil and US Treasury Bonds. The limitation of VIX is that it can only be computed for assets with large, liquid derivatives markets and hence not possible to apply VIX methods to most financial assets.

12.4 The distribution of financial returns

Return series generally are both skewed, and fat tailed and hence not normally distributed. Before we use return series in a model which assumes normal distribution we first need to check if the series is normally distributed or not. There are multiple methods to check whether the data is normally distributed or not which can be used to check the normality of the return series. We have visual methods like histogram plots or Q-Q plot and non-visual methods like Jarque Bera Test or Shapiro Wilk test. In FRM Curriculum we will discuss primarily Q-Q plot and Jarque Bera Test. Q – Q plot will be discussed in FRM Part II Book 1 Market. In this reading we will discuss Jarque Bera Test.

12.4.a Jarque-Bera Test JB Test

JB test is used to check if the returns are normally distributed. JB test is hypothesis test based measure which uses skewness and kurtosis for JB test statistics calculation.

Hypothesis statement

H_0: S = 0 and k = 3

H_A: S ≠ 0 or k ≠ 3

Where S is skewness and k is kurtosis.

Test statistics for hypothesis testing is

$$JB = (T-1)\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right)$$

Where T is the sample size.

When returns are normally distributed, the skewness is asymptotically normally distributed with a variance of 6, so that $S^2 / 6$ has a Chi square distribution. Kurtosis is asymptotically normally distributed with mean of 3 and variance of 24 and hence $(K - 3)^2/ 24$ also has a Chi squared distribution.

Decision:

JB statistics is small enough (critical value of 5.99 for significance of 5% and 9.21 for significance of 1%): The null is true which means data is normally distributed and skewness and kurtosis are 0 and 3 respectively.

JB Statistics is large (above critical value): The null is rejected, and data is not normally distributed.

12.4.b Power Laws

Power law is alternative method to check the normality of the data which study tails (fatness or thinness) if the distribution. The most important class of power law tails is which is

$P(X>x) = KX^{-\alpha}$

Where k and x are constants.

(We have power law in various topics which will elaborate this topic)

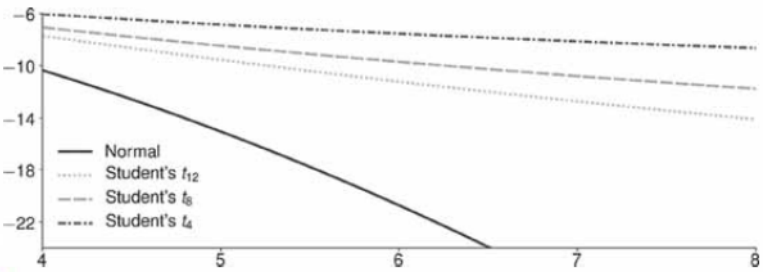


Figure 12.2 Plot of the log probability that a random variable $X > x$, where X has a mean of zero and unit standard deviation. Each curve represents $\ln Pr(X > x) = \ln(1 - F_X(x))$, which measures the chance of appearing in the extreme upper tail for $x \in (4, 8)$.

12.5 Spearman's Correlation and Kendal's T

We already discussed the Pearson's correlation coefficient which is the linear measure of correlation between two variables. The non linear form of correlation takes multiple forms. In this reading we will discuss Spearman's rank correlation and Kendal's T. Both these measures can be understood with example in better manner. We will use the following illustration in both the cases,

RETURN A	RETURN B
-10%	20%
15%	-12%
30%	8%
-20%	16%

12.5.a Spearman's Rank Correlation

Following are the steps to calculate Spearman's Rank Correlation

***Step 1 Ranking of returns*:** Start with rank return of one asset and align rank of return of another asset.

Step 2 Difference in rank: First calculate the difference in ranks D_i and take the square.

Step 3: Use this formula to calculate the correlation.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

A RETURN	B RETURN	RANK A	RANK B	DI (RANK DIFF)	DI^2
-20%	16%	1	3	-2	4
-10%	20%	2	4	-2	4
15%	-12%	3	1	2	4
30%	8%	4	2	2	4

Sum of $d_i^2 = 4 + 4 + 4 + 4 = 16$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6x16}{4(4^2 - 1)} = 0.6$$

Hence Spearman's correlation coefficient is 0.6.

12.5.b Kendal's T

Note: Please watch the Falcon Edufin video on YouTube to understand this topic in better manner.

Link: https://www.youtube.com/watch?v=OgzPdL6Vonk

A Return	**B Return**	**Rank A** (Xi)	**Rank B** (Yi)
-20%	16%	1	3
-10%	20%	2	4
15%	-12%	3	1
30%	8%	4	2

Following are the steps to calculate correlation.

Step 1: Rank returns using similar above method.

Step 2: Find concordant pairs and Discordant pairs. Concordant pair is Xi > Xj and Yi > Yj or if Xi < Xj and X Yi<Yj. Discordant pairs are the pair which is not concordant.

Concordant pairs	Discordant Pairs
(1,3)(2,4) (3,1)(4,2)	(1,3)(3,1) (1,3)(4,2) (2,4)(3,1) (2,4)(4,2)

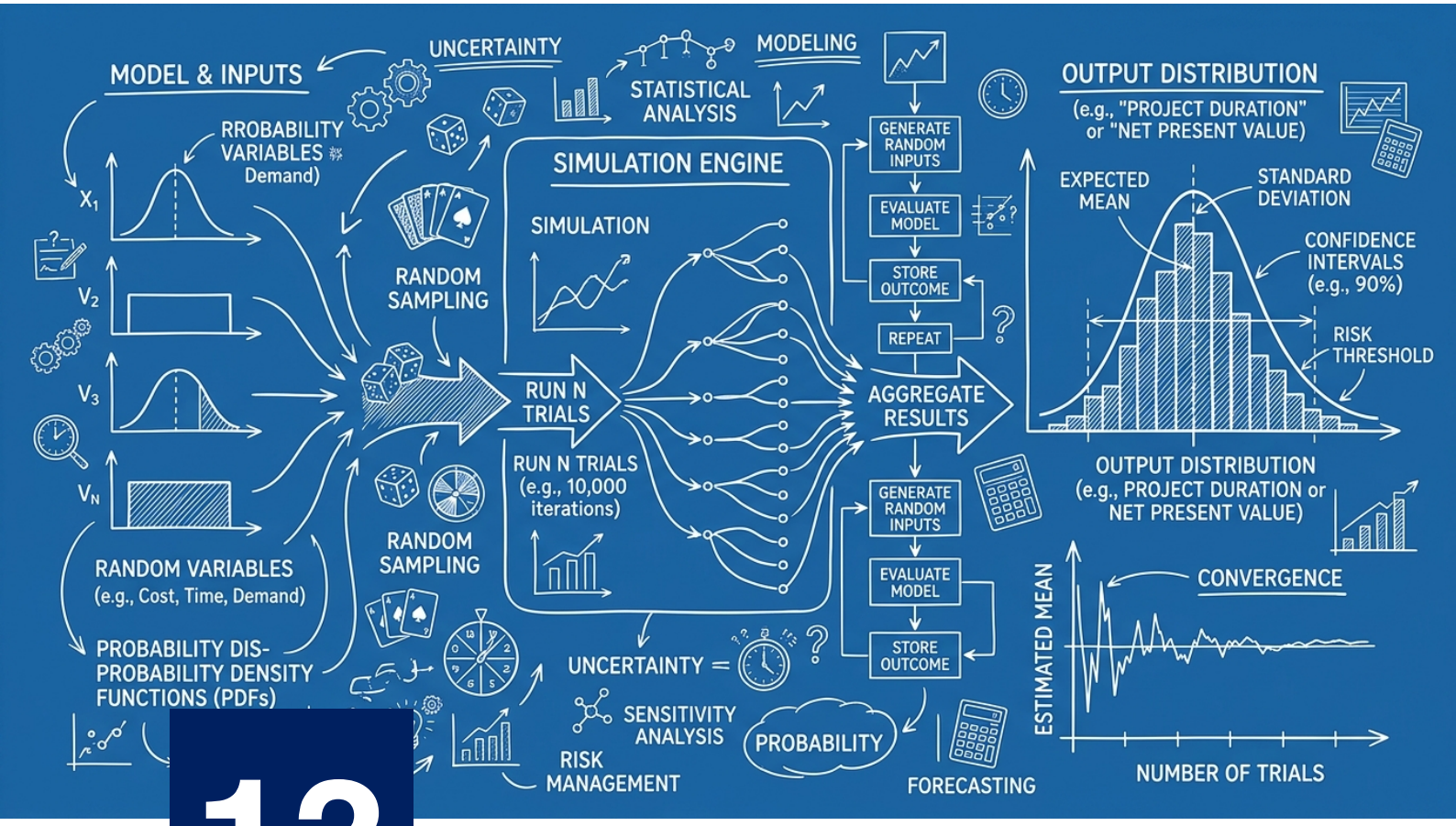
Total concordant pairs are 2 and total discordant pairs are 4.

Step 3: Calculate correlation using following formula

$$\rho = \frac{n_c - n_d}{\frac{n(n-1)}{2}} = \frac{2-4}{\frac{4(4-1)}{2}} = -0.333,$$

Hence Kendal's T correlation is -0.333.

Reading 13 Simulation and Bootstrapping



13 Simulation and Bootstrapping

SCOPE OF THIS READING

This chapter develops simulation-based techniques used in financial risk modeling. It outlines the fundamental steps in conducting a Monte Carlo simulation and examines methods to reduce sampling error, including variance reduction techniques such as antithetic and control variates. The chapter explains pseudo-random number generation and introduces the bootstrapping method, highlighting its advantages and contexts where it may be ineffective. Finally, it evaluates the limitations and practical drawbacks of simulation approaches in financial problem-solving.

13.1 Introduction: Monte Carlo Simulation

Monte Carlo simulation is the process/tool designed to approximate the expected value of the random variable using a numerical method.

Simulation experiment steps

1) Data generation X using assumed Data generation process
2) Calculate the required results or statistic like mean or standard deviation.
3) Repeat the above process
4) Evaluate the result and accuracy of simulation experiment

Data generation process starts with assuming the inputs and required distribution. Let's assume a very simple example, where we want to simulate the yearly return of a portfolio consisting of two stocks A and B invested in equal proportion. Using the historical data of return of both of these stocks, we found both the stocks follow normal distribution with following specifications

Note: Please use the following table of simulation data for better understanding.

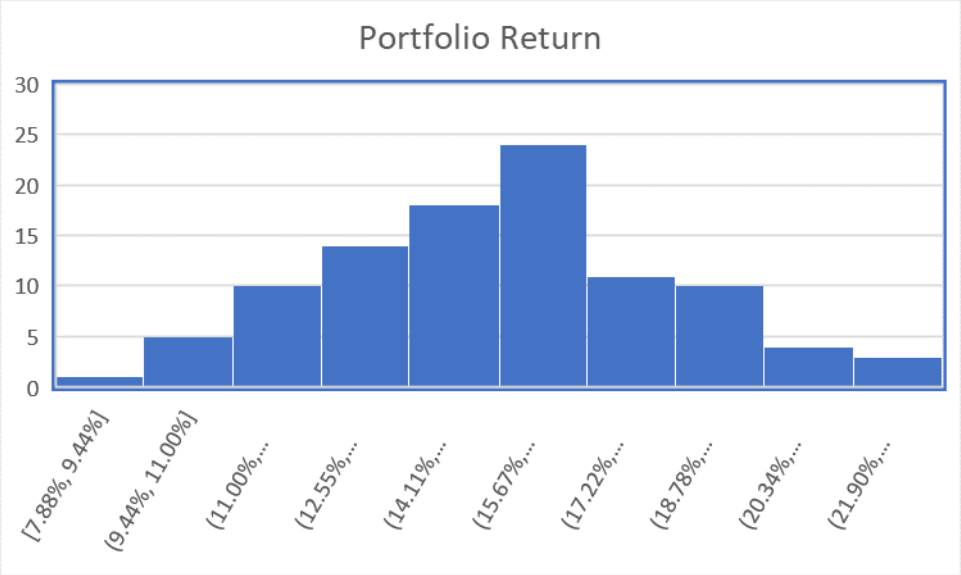
- Stock A: Mean return = 12% and Standard Deviation of return = 3%
- Stock B: Mean return = 20% and Standard Deviation of return = 5%

	Simulated data		
Trail	Stock A	Stock B	Portfolio Return
1	11.15%	19.28%	15.22%
2	16.58%	15.68%	16.13%
3	11.76%	19.52%	15.64%
4	15.15%	22.77%	18.96%
5	11.27%	25.01%	18.14%
6	8.03%	26.21%	17.12%
7	13.25%	21.41%	17.33%
8	12.61%	17.95%	15.28%
9	9.32%	24.56%	16.94%
10	12.26%	23.89%	18.08%
11	12.39%	22.00%	17.20%
12	18.11%	24.48%	21.30%
13	17.36%	29.04%	23.20%
14	11.76%	13.66%	12.71%
15	13.23%	14.34%	13.79%
16	17.47%	18.69%	18.08%
17	11.65%	23.12%	17.39%
18	11.61%	35.74%	23.67%
19	6.96%	15.96%	11.46%
20	8.95%	18.18%	13.56%
21	7.73%	13.99%	10.86%
22	10.14%	18.15%	14.14%
23	10.42%	17.95%	14.19%
24	13.25%	23.78%	18.52%
25	11.30%	21.18%	16.24%

Trial 1: We will use the specific function to generate the random data of return of each stock which is then used to evaluate the return of the portfolio. In excel you can use norm.inv() function by providing inputs of mean and standard deviation for each stock which will provide random returns for each stocks by considering the limitations put by assumed distribution. In this case we assumed the normal distribution, however, depending on the data we can assume other forms of distribution as well. Let's say we got the result as 13.5% for stock A and 17.5% for stock B. Using these results, the return of portfolio in first trail is 15.5%.

Trial 2 to n: We will repeat the same step mentioned in trail 1 n number of times which gives n results. Let us assume we simulate the data for 10,000 trails, which will give us 10,000 portfolio returns.

Please note that the returns generated for each stocks in the above process are Pseudo random and not true random. To calculate the expected return of the portfolio we use 10,000 portfolio returns calculated using above process.



Above example provides the basic idea of how the simulation can be done using the simple example. However, in practical life things are not so simple and how complex you want your simulation to be is completely your choice. In the following table I will give some examples of complex form of simulations to achieve same result (portfolio return).

Setup 1: Instead of simulating returns of stocks, we know returns are outcome of price changes and hence we can start with simulating price changes, then calculate return using price change and then calculate the portfolio return.

Setup 2: We know stock prices are outcome of volatility. So, we can start with simulate volatility instead of stock prices, then calculate stock prices using this volatility. Rest of the process is same.

Setup 3: Stock prices are outcome of various market factors. We can use regression tools to find other factors which impacts stock prices, like interest rate, companies EPS and so on. Once we get the robust regression model, we can simulate the factors first, plug it in regression equation to arrive at stock prices.

I specifically mentioned these setups just to give one lesson about simulation. Simulation process is highly flexible in its application and there is no limit how we can apply it in our models. However, we have to play in ground rules set for simulation modeling. For example, we cannot select any distribution abruptly for random value generation. These should be plausible explanation for choice of distribution. It is well known in the industry, that the distributions don't fit perfectly for real life data and hence plausible explanation is enough to use distribution.

13.2 Pseudo-Random Number Generator

Just recall the above illustrations in which we saw computer randomly provided us the returns. Obviously, these returns are within the range of stated distribution but still random. Question is how can a computer select returns randomly from the given distribution? This question is very important because computer don't have "free will" which humans have. Let say I put 3 different color balls in front of you and asks you to pick one. You can pick any random ball and the reason

is your free will. Computers don't have any free will and when we ask computers to make any choice randomly computers are incapable in it. Computers are only capable of working with mathematics. Hence computers use mathematical equation to generate random numbers designed by mathematicians. These equations are then converted into algorithm called as Pseudo Random Number Generators (PRNG). PRNG needs initial value as input value to start the algorithm and then the result of first iteration of algorithm used as an input value for second iteration. The result of PRNG is used to select random value from the distribution and this process is repeated in every trail. Because the random numbers generated by computers are not true random, hence we call it as Pseudo random number. This algorithm PRNG works behind the scenes and we will get the results which may look like a random value but are the result of PRNG.

Seed Value: We discussed in the above section that PRNG needs initial value as input. If not provided specifically, the input value can be anything (depending on algorithm) like sum of current date and time available in computer clock. The unknown inputs will produce different results every time. However, in some cases we might want to replicate the same simulation result. For example, you worked on a simulation and sent your model and simulation data to testing team. When testing team runs this model in their system computer will select different input value and they will get different results in simulation and may land you in trouble for providing wrong simulation result. In such cases we can use seed value. Seed value is the randomly selected (by user) value used as the first input in algorithm. Because algorithm is predefined and sequential, it will produce exactly same results if the first input value is same. Seed value can be anything like 123, 111, etc. Once you produce the simulation results using specific seed value, you have to provide the same seed value to your testing team which then uses the same seed value in their simulation run and will get same results. In summary seed value is mainly used to produce same simulation results and following are some scenarios in which requires seed value,

- **Where repeatability is required:** Repeatability means the production of same result for same model and input parameters. Regulators might require banks to simulate the data to calculate stressed VaR. In this case if regulator is unable to replicate the same results in simulation and stressed VaR is higher than the one quoted by bank, then banks may face the regulatory action for understating the stressed VaR. Using the seed value ensures the repeatability and hence regulator should get the same stressed VaR result as long as the model and seed value is same.
- **Cluster computing:** Assume a portfolio consisting of 100 stocks and you want to simulate the return and risk of this portfolio. Running simulation of this huge portfolio in one computer is likely to consume lot of processor load. In this case you can divide this work of simulation in two computers by dividing 50 socks in each computer to reduce process load. If no seed value is provided, then both the computers will use different random number sequence and hence results might not be comparable. In such cases same seed value is provided in both the computers to get the comparable results. Once the results are produced final results can be combined into one computer.

13.3 Improving Accuracy of simulation

The standard error of the estimated expected value depends on the variance of the simulated values and is proportional to $\frac{1}{\sqrt{b}}$ where b is the number of iterations in simulation. We know that the variance of a sum of random variables is the sum of the variance plus twice the covariances between each distinct pair of random variables.

13.4 Antithetic Variables

Antithetic variable is the simple method which is used to improve accuracy of simulation. In basic Monte Carlo simulation, we generate sample consisting of independent observations. Antithetic variables are random variables that are constructed to generate negative correlation within the values used in the simulation.

Antithetic variates add a second set of random variables that are designed to have negative correlation with the variable used in simulation. It is generated in pairs using a single uniform value. If U_1 is the variable, then antithetic variate U_2 is generated as

$U_2 = 1 - U_1$

Where U1 and U2 both are uniform random variable.

Hence by structure, correlation between both variables are negative and mapping these values through the values through the inverse CDF generates random variables that are negatively related.

Using antithetic random variable in simulation in virtually identical to running standard simulation. The difference is when generating the value used in the simulations. These random variables are then transformed to have the required distributed using the inverse CDF. Because the antithetic variables are correlated the standard error of the simulated expectations became

$$\frac{\sigma_g\sqrt{1+\rho}}{\sqrt{b}}$$

Thus, the standard error is reduced if $\rho < 0$ i.e. negative correlation.

Note: There are various approaches to apply antithetic variate technique which we will learn in practical session.

13.5 Control Variates

Control variate is the alternative method to reduce simulation error. Control variate is the random variable is correlated with error in simulation and has mean of zero. A good control variate should have two properties

- It should be inexpensive to construct from variable under simulation. If control variate is more complex and requires more time to compute, then it is better to increase number of simulations directly instead of using control variate.
- Control variate should have high correlation with statistic in simulation.

13.6 Limitations of Simulations

The challenge in using simulation to approximate moments is the specification of the DGP. If the DGP does not adequately describe the observed data, then the approximation of the moment may not be reliable. Misspecification in DGP can occur due to

- Choice of distribution is incorrect
- Using incorrect parameters estimates to simulate the data.

One more important consideration is the computational cost. In modern computers single simulations (basic level) won't take more than a minute but for running complex simulation in large number can be time consuming.

13.7 Bootstrapping

In the earlier section we learned the process to generate the data using simulation. Bootstrapping is the alternative method to generate the data. Simulation and bootstrapping are both generates data using historical data sets, but the approach is different. In simulation we used the data to get parameters which are then used in the form of distribution to generate the data. In contrast, bootstrapping directly uses the historical data to simulate sample with similar characteristics. Bootstrapping does not require any assumption relating to distribution due to use of direct data in the data generation process. There are two classes of bootstraps using in the risk management techniques,

iid Bootstrap: It is simple because samples are created by drawing with replacement from the observed data. Assume the data set of n and you want to generate the data simulation data with m observations. In iid bootstrap data is generated directly by random sampling with replacement from n observations. Iid bootstrap is applicable when observations are independent across time.

Circular block bootstrap: In some cases, more sophisticated bootstrap method is required. One such method is circular block bootstrap or CBB. This method is similar to iid bootstrap with only difference being instead of directly sampling from observed data we sample size of q with replacement.

Let's assume the following data

A B C D E F G H I J K L

We can create block of say 2 elements

(A,B) (B,C) (C,D) …. (K,L) (L,A)

In circular block instead of sampling data randomly, we sample block of data and repeat the process.

To generate a data using the CBB method,

- Select the block size q
- Select the first block(randomly) from the created blocks to bootstrap sample.
- Repeat the step 2
- If the bootstrapped sample has more than m elements drop values from the end of the bootstrap samples until the sample size is m.

The choice of block size q should be large enough to capture the dependence in the data, although not so large as to leave too few blocks. The rule of thumb is to use a block size equal to the square root of the sample size $\sqrt{n}$.

Bootstrapping methods can be ineffective in following cases

- **Change in market conditions:** Using historical data is only useful if the current market conditions are same. If there is change in current market conditions say change in volatility regime, then using historical data will produce unreliable results.
- **Fundamental changes**: If the fundamentals of markets in the history were different compared to current market conditions. For example, formula used to calculate GDP was

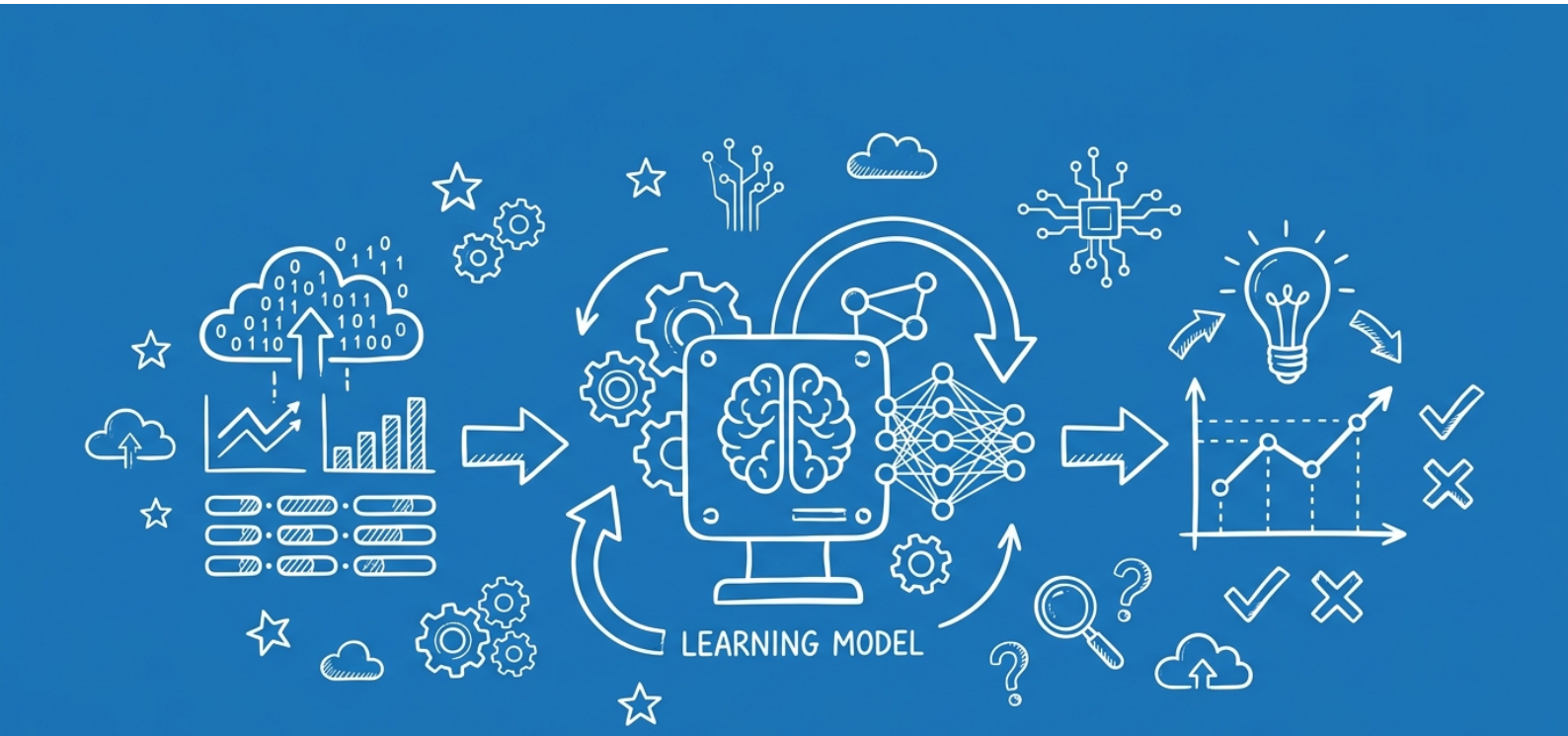
different in history compared to current formula. Using old GDP data in current simulation might not produce accurate results.

13.8 Disadvantages of Simulation

Simulation is not always dependable and depends on lots of assumptions relating to models or distributions. Hence, lot of scholars recommends if you have choice in between the close form equation and simulation, one should prefer closed form equation. For example, to calculate the option price we have BSM formula which is closed form equation. However, we can also use simulation to calculate option price by simulating stock prices at maturity and then calculating payoffs. The expected payoff of simulation is option price. However, it is recommended to prefer BSM formula instead simulation approach because of the following disadvantages of simulation processes.

- **Unreliable DGP:** Data generating process depends on model specification and distribution assumption. A slight variation in choice of distribution or model specification might result into different simulation outputs.
- **Simulation computation costs:** Simulation is computer dependent and requires intensive processing power. With modern computers we can get the results for basic simulation in few seconds which was not possible few decades back. However, for more complex simulations computational cost can be extremely high. This might be in the form of hardware cost or time to compute.

Reading 14 Machine Learning Methods



14 Machine Learning Methods

SCOPE OF THIS READING

This chapter contrasts machine learning techniques with classical econometrics from both philosophical and practical perspectives. It explains the roles of training, validation, and test datasets, and analyzes the implications of underfitting and overfitting along with appropriate remedies. The chapter introduces dimensionality reduction using principal components analysis and clustering through the k-means algorithm. It also provides an overview of natural language processing applications and differentiates among supervised, unsupervised, and reinforcement learning frameworks, explaining how reinforcement learning operates in sequential decision-making contexts.

14.0 Concept of Machine Learning

Machine learning is a type of computer science that uses algorithms to teach computers how to learn from data without explicit programming. It is part of artificial intelligence (AI) and is used to make models that can predict outcomes.
Machine learning is about finding patterns and trends in data by using labeled data to train a machine learning model. Labeled data is data which has the right output attached. For example, if the data is pictures of cats and dogs, then the data would have "cat" or "dog" labels. The machine learning model learns to find patterns and trends in the data from these labels. After the model is trained, it can predict on new data. For example, if the model is trained on cat and dog pictures, it can tell if a picture has a cat or dog. This is called supervised learning, where the model learns from labeled data and predicts on new data.

Machine learning use cases:

- Image Recognition: A machine learning application that can identify objects, people, scenes, and activities in images.
- Natural Language Processing: A machine learning application that can understand, analyze, and generate human language.
- Recommender Systems: A machine learning application that can suggest items to users based on their past interactions and preferences.
- Credit Risk Analysis: A machine learning application that can identify and assess the risks associated with giving credit to customers.
- Fraud Detection: A machine learning application that can detect suspicious activity and determine whether it is likely to be fraudulent or not.
- Autonomous Driving: A machine learning application that can make decisions and take actions in order to safely navigate a vehicle.
- Robotic Process Automation: A machine learning application that can automate mundane or repetitive tasks.
- Medical Diagnosis: A machine learning application that can diagnose medical conditions by analyzing patient data.

Machine learning helps to enhance decision-making and automation in finance. Some examples of its applications are credit scoring, stock market forecasts, fraud detection, portfolio optimization, and algorithmic trading. Machine learning algorithms can find patterns in data and predict future results. They can spot anomalies and reveal insights that are hard to find. Also, machine learning can automate tasks such as portfolio rebalancing, risk management, and customer segmentation. With machine learning, financial institutions can make better decisions and improve their operations.

14.1Types of machine learning

Machine learning methodologies can be categorized as follows:

Supervised Machine learning: Supervised machine learning is a type of machine learning where a model learns from data that has labels to make predictions on new, unseen data. For example, a supervised machine learning model can be used to predict the price of a house based on data points with labels such as square footage, number of bedrooms, and location. The model learns from a set of labeled data and can then be used to make predictions on new data points.

Supervised machine learning has been used in finance for various tasks, such as predicting stock prices, identifying customer churn, and forecasting credit risk. To use supervised machine learning for these tasks, data sets with historical financial information must be collected and labeled. Then, the data is input into a machine learning algorithm which is trained to find patterns in the data and make predictions. These predictions can then be used to support financial decisions and strategies. Furthermore, supervised machine learning can be used to detect fraudulent activity, such as insider trading.

Unsupervised Machine Learning: Unsupervised machine learning is a form of artificial intelligence (AI) that searches for hidden patterns in a data set without the guidance of a human supervisor. It is used to make inferences from datasets that consist of input data without labeled outcomes. The aim is to find structure in the data, which can then be used to forecast future behavior.

Unsupervised machine learning algorithms are used to discover patterns and relationships in data sets that would otherwise be difficult to detect. These algorithms can be used to group data points into separate clusters or groups based on their similarities. They can also be used to identify anomalies and outliers, and to extract useful features from data.

Unsupervised machine learning algorithms are generally classified into two types: clustering and dimensionality reduction. Clustering algorithms arrange data points into different groups or clusters according to their likeness. Dimensionality reduction algorithms decrease the number of variables or features in a data set without losing significant information.

Unsupervised machine learning can be used for various applications, such as anomaly detection, data visualization, market segmentation, recommendation systems, and natural language processing. It can also be used to find patterns in data that are not easily noticeable, such as fraud or outliers.

Reinforcement Learning: Reinforcement learning (RL) is a type of machine learning where agents learn how to act in an environment so as to increase a total reward. It is an area of artificial intelligence where an agent learns to interact with its environment to improve its performance.

The essence of reinforcement learning is that an agent learns from its environment by trying out different actions and getting feedback for each action. The agent gets a reward for each action, and it learns to optimize this reward by choosing the best action for each state. The agent follows some rules, called a policy, to decide the best action in each situation. This policy is then revised as the agent learns from its environment, allowing it to enhance its performance over time.

Reinforcement learning is useful for various machine learning tasks, such as robotics, natural language processing, adaptive control, and game playing. It can also be applied to solve complex optimization problems. Therefore, it has become a valuable tool in the creation of autonomous systems.

Key points to remember:

- ***Unsupervised machine learning is not used to generate predictions.*** It is used to characterize a dataset and learn its structure. For example, Unsupervised ML can be used for anomaly detection where bank is trying to find features of transactions that might be suspicious and worthwhile of further investigation.
- Reinforcement learning is useful in risk management. For example: to determine the optimal way to buy or sell a large block of shares.

14.2 Data Preparation

Data preparation is very important for effective machine learning model and prediction. Following are the steps of data preparation.

1. Data Cleaning: This is the process of identifying and correcting or removing corrupt or inaccurate records from a dataset. This is especially important for machine learning models as any anomalies or outliers in the data may produce unexpected and inaccurate results.
2. Feature Engineering: This is the process of transforming raw data into features that are more meaningful and useful for machine learning algorithms. This includes feature selection, feature scaling, feature extraction, etc.
3. Data Splitting: This is the process of splitting the data into train and test sets for model training and evaluation. The train set is used to train the model, while the test set is used to evaluate the performance of the model.
4. Data Normalization: This is the process of scaling the data so that it follows a normal distribution. This is important as it helps ensure that the data is in a consistent format and that all features are treated equally by the model.
5. Data Augmentation: This is the process of adding additional data to the original dataset to create a more robust model. This may include adding additional features or generating synthetic data.

Data cleaning

Data cleaning in machine learning is the process of preparing data for analysis. It involves identifying and removing errors, outliers, inconsistencies, and duplicate data. Data cleaning also involves transforming the data into a format that is suitable for the machine learning model being used. This includes the encoding of categorical data, normalizing numerical data, and creating additional features from the dataset. Data cleaning also involves filling in missing data, such as by using imputation techniques.

Reasons for data cleaning

- **Inconsistent recording:** All the data should be recorded in the same way. For example, in the date column, dates are recorded in different formats which will create reading difficulty for ML model.
- **Unwanted observation:** Observations not relevant to the project should be dropped from the data. Keeping unwanted observations can impact results as well as computational time.
- **Duplicate observations:** Should be removed to avoid biases.
- **Outliers:** Outliers may affect the standard deviation from the mean which might affect the final results. Outliers should be dealt with correctly (by dropping or scaling, discussed in the following section)
- **Missing data:** This is a common problem. If there are very few observations that are missing in the data set then it can be dropped. Alternatively, missing observation can be replaced by mean or median of the non missing observation. There are other approaches of replacing missing data which are more complicated (like average of adjacent observations).

Data scaling (Standardization and normalization)

Scaling is an important step in machine learning because it helps to normalize the data. This means that all of the features in the dataset are on the same scale, which helps the model to learn more effectively. It also helps to reduce the influence of outliers, which can have a significant impact on the model's performance. Scaling can also help to improve the accuracy of the model since it allows the model to capture patterns in the data.

Standardization is the process of rescaling a variable so that it has a mean of zero and a standard deviation of one. This is usually done by subtracting the mean from each value and then dividing it by the standard deviation. Standardization is useful for data sets that have different scales and units of measurement.

$$X_{ij} = \frac{X_{ij} - \hat{\mu}}{\hat{\sigma}_i}$$

Normalization is the process of rescaling a variable so that it has a range of values between 0 and 1. This is usually done by dividing each value by the maximum value in the data set. Normalization is useful for data sets that have different scales and units of measurement.

$$X_{ij} = \frac{X_{ij} - X_{ij,min}}{X_{ij,min} - X_{ij,max}}$$

Regardless of which method we use, all the inputs must be rescaled. However, recasling is not necessary for prediction. The choice of method standardization or normalization depends upon the nature of data. Standardization is preferred when the data covers the wide scope, including outliers. Normalization would squeez the data points into tight range which may not have the similar characteristics like the original data.

14.3 Principle component analysis

Principal Component Analysis (PCA) is a technique used in unsupervised machine learning to reduce the number of features in a dataset while retaining as much of its variance as possible. PCA works by transforming the dataset into a set of orthogonal components which are uncorrelated and capture most of the variance within the dataset. This reduces the dimensionality of the dataset, making it easier to work with and more efficient to process. PCA can also be used to identify patterns and correlations within the dataset which can be used to gain insights into the data. Following are the uses of PCA

1. Dimensionality Reduction: PCA is widely used for dimensionality reduction in machine learning applications. It helps to reduce the number of dimensions or features of a dataset by removing redundant information and preserving the most important features of the dataset.
2. Feature Extraction: PCA is also used in machine learning as a feature extraction technique. It helps to identify the most important features in a dataset and extract them for further analysis.
3. Visualization: PCA is often used for data visualization in machine learning. PCA can be used to reduce the number of dimensions in a dataset and create a 2-dimensional or 3-dimensional visual representation of the data. This can help to identify clusters and patterns in the data.
4. Outlier Detection: PCA can be used to detect outliers in a dataset by identifying points that have a high distance from the centroid.

5. Noise Filtering: PCA can be used to remove noise from a dataset by identifying and removing components with low variance.

PCA has been used in finance to analyze stock returns, identify patterns in financial time series, and to reduce the dimensionality of financial data. It can also be used to identify which stocks are closely related and to construct portfolios with a diversified risk profile. PCA can also be used to detect outliers and to highlight clusters of stocks with similar characteristics. PCA can also be used to construct portfolios that have the highest expected return given a certain level of risk.

PCA Application (from GARP Book)

A typical application of PCA is to reduce a set of yield-curve movements to a small number of explanatory variables or components. Suppose, for instance, that we have ten years' worth of daily movements in interest rates with one-month, three-months, six-months, one-year, three-years, five-years, ten-years, and 30-years maturity. The aim in PCA is to find a small number of uncorrelated variables that describe the movements. Specifically, the observed movements should, to a good approximation, be a linear combination of the new variables.

For yield-curve movements, the most important explanatory variable is a parallel shift where all interest rates move in the same direction by approximately the same amount. The second-most important explanatory variable is a "twist," where short rates move in one direction and long rates move in another direction.

Following table shows the principal components constructed from monthly movements in seven Treasury rates between January 2012 to December 2021 (120 data points). To explain the movements fully, seven components are necessary. However, when the actual movements are expressed as a linear combination of the components, the first (approximately parallel shift) component explains most of the variation (73.3%), and the first three components explain more than 99% of the variation. This is because there is a high degree of correlation between the yield movements, and the bulk of the information contained in them can be captured by a small number of explanatory variables.

	Principal component For US Treasury bill and bond series						
Series	1	2	3	4	5	6	7
USTB1M	0.41	0.264	0.3	-0.568	-0.151	0.499	-0.279
USTB3M	0.415	0.253	0.227	-0.194	0.59	-0.492	0.289
USTB6M	0.42	0.234	0.093	0.258	-0.722	-0.41	0.069
USTB1Y	0.424	0.201	-0.1	0.699	0.297	0.422	-0.122
USTB5Y	0.405	-0.226	-0.757	-0.269	-0.062	0.114	0.351
USTB10Y	0.31	-0.541	-0.05	-0.016	0.107	-0.319	-0.704
USTB20Y	0.21	-0.654	0.514	0.108	-0.066	0.218	0.447

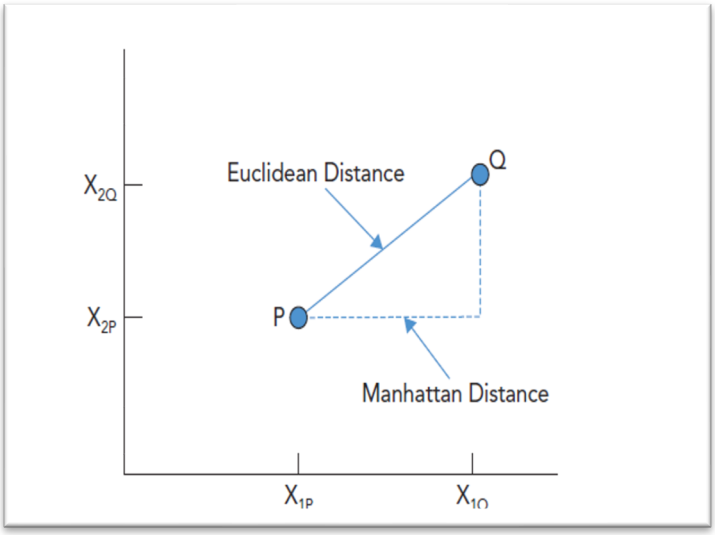
14.4 The K-Means Clustering Algorithm

K-means is an unsupervised machine-learning algorithm used to group data into clusters based on similarities. It is the most commonly used clustering algorithm and works by finding k clusters in the data, where k is an integer specified by the user. The algorithm works by randomly initializing k centroids, then assigning each data point to the closest centroid based on Euclidean distance. The centroids are then moved to the mean of the points assigned to each cluster, and the process is repeated until the centroids no longer move. The result is a set of k clusters, with each cluster represented by its centroid.

1. Select the number of clusters (k) on random basis.
2. Select random k points as centroids.
3. Assign each data point to the nearest centroid based on distance (Euclidean or Manhattan Distance).
4. Compute and place the new centroid of each cluster.
5. Reassign each data point to the new nearest centroid.
6. Repeat Steps 4 and 5 until the centroids no longer move.

Euclidean Distance (Direct route distance): Euclidean distance is used for clustering data points. It works by measuring the Euclidean distance between each data point and the cluster center and then assigning points to the nearest cluster. The algorithm continues until the clusters converge, meaning that all points within the same cluster are closer to the center than to any other cluster. This type of clustering is often used in situations where the data points are in a multi-dimensional space. It is a useful tool for clustering data points based on their similarity and can be used to identify patterns in data.

Manhattan Distance: Manhattan Distance uses the Manhattan distance metric to cluster data points. The Manhattan distance is the sum of the absolute differences between two points on a plane. It is used to cluster data points that are close together into distinct groups. This method is useful for applications such as market segmentation, customer segmentation, and image segmentation. This clustering method is easy to implement, and can be used to quickly and accurately identify clusters.



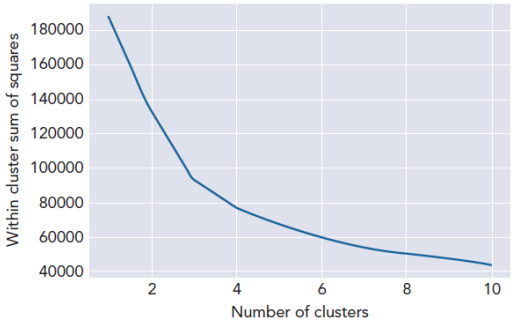
Performance Measurement for K-Means

For K-Means, the most common performance measure is the Sum of Squared Errors (SSE) or known as inertia, which measures the sum of the squared distances between each point and the centroid of its assigned cluster. The lower the SSE, the better the performance of the clustering algorithm. Additionally, the silhouette coefficient is also used to measure the performance of K-Means. The silhouette coefficient measures the similarity of a data point to its own cluster compared to other clusters. The higher the silhouette coefficient, the better the performance of the clustering algorithm.

Selection of K

Unlike R^2 which never falls when the explanatory variable is added, the inertia will never rise as the number of centroids increases. The maximum possible value of K is the total number of data points and in this case, each observation will form its own clusters. When the cluster K=n, the inertia is equal to zero. The choice of K should be practical.

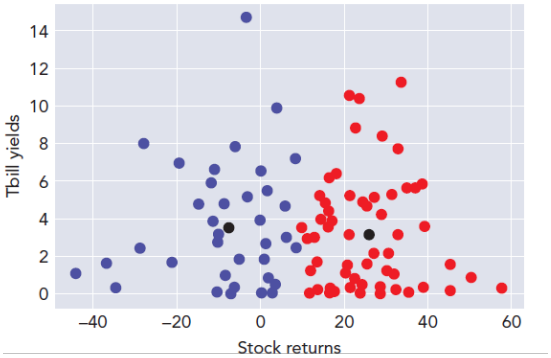
The scree plot provides the value of inertia for different values of K. The scree plot can be utilized to determine the number of components to use in PCA. We would examine the graph to determine where there is an obvious point at which inertia starts to decline more slowly as k is further increased, which is then choose as the optimal number of centroids. Please check the following scree plot from GARP book.



The choice of K from scree plot is the slight elbow shape point visible in between 2 and 4, indicating the value might be optimal.

An alternative way to choose K is the silhouette coefficient. This compares the distance of each observation from other points in its own cluster with its distance from points in the closes other cluster. ***The Best value of K is the one which gives the highest silhouette score***.

Apart from selecting a priori number for clusters, the other disadvantage of the technique is that because it is based on distances from a centroid, it tends to produce spherical clusters.



14.5 Machine learning vs Traditional Econometrics (Linear regression and Time Series Forecasting)

Machine learning offers advantages over the traditional linear econometric approaches in the forecasting/prediction.

- Machine learning works well even if there is constrained financial theory is available to guide the choice of variable to include in a model or whether e researcher in unsure about the linear nonlinear method is more suitable for forecasting.
- Machine learning can capture complex forms of relation between variables. Like when the two variables are correlated, in traditional model, researcher needs to specifically structure model to capture such correlation however, Machine learning methods will consider the impact of such correlation on overall model.

The model construction approaches are different in traditional modeling and machine learning. Also, the methods to evaluate the model efficiency are different in both. Machine learning does not apply the methods like statistical significance, goodness of fit and error term diagnostics testing to evaluate the model which is used in traditional models. Machine learning instead focuses on the accuracy of prediction.

Machine learning does not require the data distribution assumption whereas the traditional models are heavily dependent on the normal distribution assumption of the data.

Although there are differences in ML and traditional models, we can say, standard regression specification is the special case of advanced machine learning like neural network.

Machine learning methods are developed by engineers and not by statisticians. Hence we see the difference in nomenclature used in machine learning. For example, the variables of conventional econometrics are called as inputs or features in machine learning, similarly, dependent variable is called as output or target.

14.6 Overfitting and underfitting

Overfitting

Overfitting in machine learning refers to a model that has been excessively trained and is no longer able to generalize to unseen data. It occurs when a model is excessively complex, such as having too many parameters relative to the number of training samples. The model learns the training data too well and memorizes it, but fails to generalize to new data. As a result, the model performs well on the training data, but does not perform well on test data.

Overfitting in machine learning occurs when a model learns the details and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the model will have a good accuracy on the training data but will not be able to generalize well to new data. In other words, it has memorized the training data, instead of learning the real underlying patterns. One example of overfitting in machine learning is when a model is trained on a dataset that is too small. This can cause the model to learn patterns that are specific to the dataset, but may not generalize well to new, unseen data.

Key points - overfitted model

- Performs poorly on new data.
- Captures excessive random noise in the training set.
- False impression of an excellent specification.
- Overfitting is more common and problematic in machine learning compared to traditional econometrics.

Underfitting

Underfitting in machine learning is when a model fails to capture the underlying pattern of the data and is unable to make accurate predictions. This is usually caused by a model that is too simple or by using insufficient data to train the model. This can lead to high bias and low variance, resulting in an inaccurate model. An example of underfitting would be a linear regression model that has been trained on a nonlinear dataset. The model would be unable to capture the nonlinear relationship between the inputs and the outputs and would thus underfit the data and not make accurate predictions.

Key points – Underfitted model

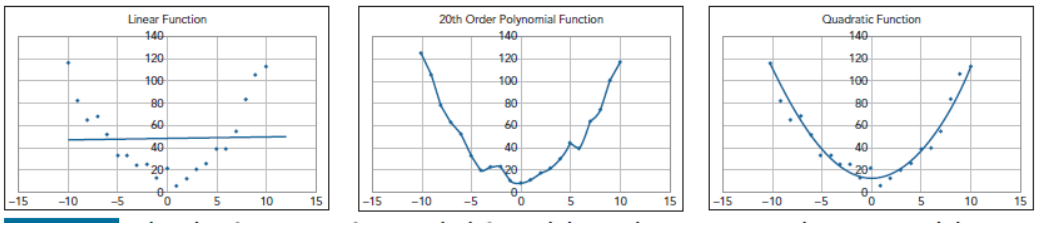
- Underfitting can be caused by number or quality of inputs is insufficient or steps not taken to avoid underfitting.
- Failure to include relevant interaction terms can result into underfitting

The choice of the size of the ML model determines the underfitting, overfitting or proper fitting of the data. The choice involved are termed as bias variance tradeoff. We discussed this in Reading on Regression.

Key points

- Underfitted model – Higher bias with low variance
- Overfitted model – Lower bias but higher variance

Following from GARP curriculum book shows three graphs.



14.7 Sampling and Splitting and Preparation

Training, Validation and Test data

In traditional econometric models, we collect the data and divide it into two data sets known as the training data set and test data set. Let's assume we have a total of 100 observations. We can divide these observations into two parts 80% i.e. 800 randomly selected observations as selected training data and the remaining 200 as testing data. Please note, the random selection of data points is only applicable for cross sectional data and not applicable for time series data. We will discuss data splitting of time series in following section. There is no fixed rule for the selection of training and testing data proportion, but the most commonly used proportion for training data is around 80% of total observations. *This is not GARPs view on data split percentage.*

- **Training set**: Also referred as within sample is the data used to find the model parameters using which the models are trained and selected.
- **Testing set**: Also referred as out of sample data used to test the selected model. In testing we check the prediction power of model on test data compared to training data. As we learned in previous section, if model performs well on training data but fails to perform similarly on the testing data then there might be overfitting problem.

Data splitting in Machine learning models:

In machine learning models data is split into three parts training data, validation data and test data.

- **Training set:** Same as discussed in training data for traditional models. However, the use of this data in ML is limited to model building/training.
- **Validation set:** Assume we have three competing models from previous step. To compare these models, we need to check which model best generalizes the validation data. Once we get the final model, this data is not independent which can be used for testing.
- **Testing set:** Also known as hold out sample is used to test the model's performance compared to training data.

As we discussed, there is no fixed rule for data split percentages. As per GARP around 2/3rd data should be reserved for training and the remaining data should be divided equally as validation and testing data set. Also, we need to keep this in mind, the training data should be enough to train the model. When we have enough data then split rule is not very important because there will be enough data to train models. However, if the training set is too small, this can lead to biases in parameter estimation and small validation set will lead to model evaluation inaccuracies.

Data splitting in Time series data: For the time series data, data point selection is not random, it should be in sequence. Assume we have 364 days data, the first part (say first 200 days) should be training data, second part (next 82 days) as validation set and last part is test data set. This provides the advantage of testing data on the most recent observations.

Cross Validation Searches

When the data set is limited, the cross validation is deployed for more efficient use of data. Cross validation involves combining the training and validation data into single sample and holding test data separate. Then combined data are split into two, with estimation being performed repeatedly and one of the subsample left out each time.

Cross validation searches are a method of evaluating a machine learning model's performance by splitting the original data into multiple sets and testing each set against the model. This is done to help prevent overfitting, which occurs when a model has been trained too extensively on the same data set and is unable to generalize to new data. Cross-validation searches allow for a more robust evaluation of the model's performance by testing it on unseen data. The most common type of cross-validation search is k-fold cross-validation, which splits the data into k equal parts and tests each part once, rotating through each part to ensure that all parts are tested. This is done to ensure that the model is tested on data that it has not seen before, which allows for a more accurate evaluation of its performance.

In K-fold-cross validation, the data is split into k samples, with test data excluded. It is common to choose k=5 or 10. Then the training data would be split into 5 equally sized, randomly selected sub sample. The first estimation would use samples k1 to k4 with k5 left out. Then next estimation will select the four samples but this time left out sample is other that k5. At the end, k validation samples that can be averaged to determine the models performance.

A large value of k will imply an increased training sample which might be valuable if overall observations are low. When k =n, only one observation is left out, this method is known as leave-one-out cross validation.

14.8 Reinforcement Learning

Reinforcement learning is concerned with policy development for a series of decisions to maximize a reward. Watch documentary AlphaGo where computer program (using reinforcement learning) is developed which defeats the professional human Go player. The algorithm learns by playing against itself many times and using a systematic trial and error approach. This can be used in stock trading, hedging techniques in risk management field. Please note, the reward for machines are not similar to reward for humans. Reward is programmed explicitly in the algorithm which machine tries to achieve.

Reinforcement learning works in terms of states, action and rewards. The state is defined environment, action is decision taken and the aim is to take the decision to maximize reward. A discount rate may be used to determine the value of the total subsequent rewards.

On each trail, it is necessary to determine the actions taken for each states encountered. If the algorithm goes for best action discovered so far, it may not be able to experiment with new actions. To overcome this, the algorithm chooses between strategies that are referred to as exploration and exploitation. Which means algorithm has to decide between the best choice so far or trying new action using preassigned probabilities. The probability of exploitation increases as more trains are concluded so that algorithm learns more about best strategies.

There are two approaches to seek reward. First is known as Monte Carlo method in which algorithm takes the action in specific environment and total subsequent rewards prove to be R. Alternative, method is known as temporal difference learning. This looks only one decision ahead and assumes that the best strategy identified so far is made from that point onward.

14.9 Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence that enables machines to understand and process natural language input. It is used to analyze and interpret written or spoken text as well as to generate meaningful responses. NLP uses Machine Learning to analyze text and find patterns in the data.

Machine Learning (ML) is a method of data analysis that automates the process of recognizing patterns in large amounts of data. By providing a set of algorithms and techniques, ML can be used to analyze text and determine its meaning.

NLP using ML starts with tokenization. This is the process of breaking down a sequence of text into smaller pieces called tokens. The tokens are then identified, classified, and tagged so that the algorithm can understand their meaning. After this, the text is parsed, which is the process of analyzing the text to identify the parts of speech, such as nouns, verbs, adjectives, and adverbs.

Once the text is parsed, it can then be analyzed to identify the context of the text. Context is the underlying meaning of the text and it is used to interpret the text and determine the intended meaning.

Finally, the text is used to generate responses. This is done by using a set of algorithms.

Uses of NLP:

- Recognition of specific words to determine the purpose of a message (used in automated caller system)
- Categorization of a particular piece of text. Like Google provides the result based on search queries.
- Determine the sentiment of a statement. Marketing team using NLP to determine from social media comments to check new products response.

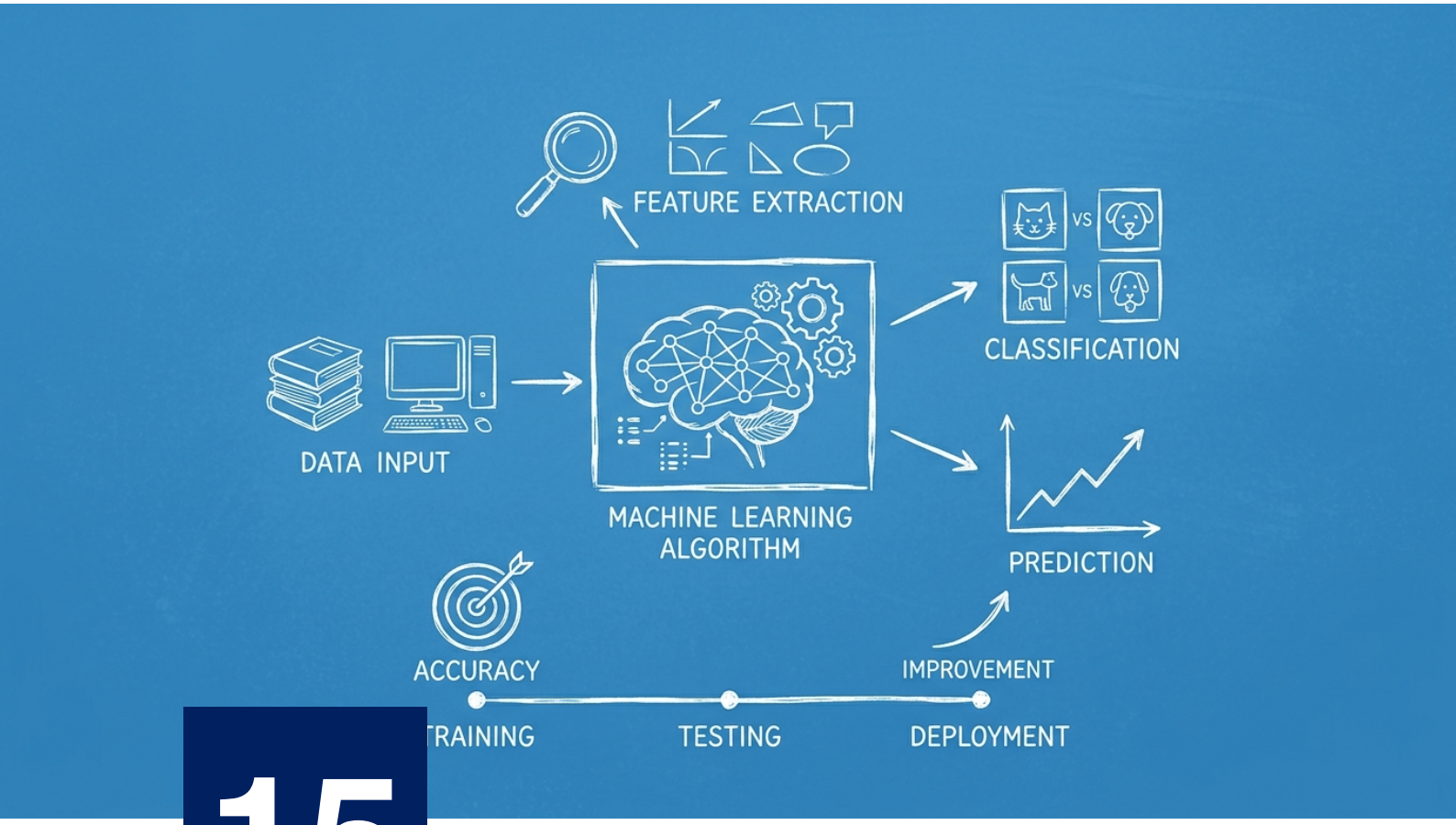
Steps in NLP process

- Capturing the language
- Pre processing the text and
- Analyzing it for a particular purpose

Preprocessing requires several intermediate steps to ensure the accuracy of analyzed text

- Tokenize the passage: Separating the piece into words, usually ignoring any punctuation, spacing, special symbols and so forth.
- Stop word removal: Stop words are those who have no information value such as a, the, also etc.
- Replace words with their stems (stemming): Where words such as disappointing and disappointed would be replaced with disappoint.
- Replace words with their lemmas: This process is sometimes known as lemmatization, where words such as good and better are replaced with good.
- Consider ”n-grams”, These are groups of words with specific meaning when placed together that need to be considered as a whole rather than individually.

Reading 15 Machine Learning and Prediction



15 Machine Learning and Prediction

SCOPE OF THIS READING

This chapter introduces core supervised learning techniques used for prediction and classification in risk analytics. It explains the roles of linear and logistic regression, including encoding of categorical variables and the rationale for regularization, distinguishing between ridge and LASSO methods. The chapter develops tree-based models, including decision tree construction and interpretation, and explains how ensemble methods improve predictive performance. It outlines classification approaches such as k-nearest neighbors and support vector machines and describes the structure and training of neural networks. Finally, it evaluates model performance using confusion matrices, particularly for logistic regression and neural network models

15.1 Dealing with Categorical Variables

Categorical variables are variables that contain label values rather than numeric values. These are variables that describe a 'characteristic' of an observation. Examples of categorical variables include race, sex, age group, and educational level. In machine learning, categorical variables are often encoded as integers or one-hot vectors.

1. One-hot encoding: This involves creating a dummy variable for each distinct category of the categorical variable. This is one of the most widely used methods in dealing with categorical variables.
2. Label encoding: Label encoding involves assigning a numerical value to each category of the categorical variable.
3. Frequency encoding: Frequency encoding involves replacing the categorical variable with the frequency of its occurrence.
4. Binary encoding: Binary encoding involves replacing the categorical variable with binary digits which represent each of the categories.
5. Target encoding: Target encoding involves replacing the categories of the categorical variable with the mean of the target variable.

Assume we have four categories of candidates pursuing FRM. Finance professionals, non finance working professionals, full time students, neither students and nor working professionals. The first approach can be to allocate the number like 0,1,2 etc for each category. However, this is ordering 0<1, but actual categories does not have any ordering. Hence we use 0-1 dummy variable as

- Finance professional: 1000
- Non finance professional: 0100
- Full time student: 0010
- Neither students and nor working professional: 0001

This is known as one-hot encoding. There may be dummy variable trap if there is an intercept and dummy variable in the model, which would mean that there is no unique best fit solution. This can be solved using regularization approaches that are discussed later are a way of handling this and a unique solution where the coefficients of the dummy variable are as small in magnitude as possible to create.

If there is natural ordering in categories then we could use dummy variables such as 1,2,3 etc.

15.2 Regularization

Regularization is a technique used in machine learning to prevent overfitting. It does this by penalizing overly complex models, reducing their complexity and thereby reducing their variance. This can be done by adding a regularization term to the cost function which penalizes weights that are too large. This forces the model to use smaller weights, making it simpler and less prone to overfitting. Regularization can also be applied by using techniques such as cross-validation, L1 and L2 regularization, and early stopping.

Two common regularization techniques are ridge regression and least absolute shrinkage and selection operator (LASSO). Both work by adding a penalty term to objective function that is being minimized.

Ridge Regression

Ridge regression is a form of regularized linear regression that is used to create models that predict quantitative values. It is similar to linear regression, but it adds a regularization term to the cost function, which prevents overfitting and makes the model more generalizable. The regularization term is the sum of the squares of the coefficients of the model and is multiplied by a tuning parameter, lambda. The larger the value of lambda, the higher the regularization and the more generalizable the model.

The cost function of ridge regression can be written as:

$C = \sum i(yi - (\beta 0 + \beta 1xi1 + \beta 2xi2 + \ldots + \beta pxip))2 + \lambda \sum j\beta j2$ (not an important equation to remember)

Where yi is the observed value, xij is the jth predictor, βj is the coefficient of the jth predictor, and λ is the tuning parameter.

The tuning parameter, lambda, is used to control the amount of regularization used in the model. A higher lambda value will lead to more regularization, while a lower lambda value will lead to less regularization.

The goal of ridge regression is to minimize the cost function by finding the optimal values of the standard error.

LASSO

Lasso is similar to ridge regression but the penalty in ridge is squared however, in Lasso its absolute. Due to the second- and first-order structure of the penalty components, ridge regression and LASSO are commonly referred to as L2 and L1 regularisation, respectively. There is a significant distinction between them. The size of the b parameters is often reduced via ridge regression (L2), bringing them closer to zero but not quite there. As a result, the model is made simpler and it is prevented that two correlated variables have one assigned with a big positive coefficient and the other with a large negative coefficient. In contrast, LASSO (L1) zeroes out some of the less significant b estimations. Depending on the circumstance and if removing or reducing extreme parameter estimations is the goal, one strategy may be preferred to the other.

Elastic net: Is the combination of above two methods, were the loss function contains both squared and absolute value function of the parameters. Combines both the methods, the benefit is reducing the magnitude of some parameters and removing some unimportant ones entirely.

15.3 Logistic Regression

Logistic regression is a statistical method used for predictive analysis. It is a supervised learning algorithm used to classify data into two categories (binary classification). Logistic regression predicts and explains the likelihood of a certain event based on a set of independent variables.

Logistic regression works by using a linear model to estimate the probability of an event occurring. A linear model is a mathematical equation that describes a linear relationship between two or more variables. The linear model used in logistic regression is often referred to as the logit function.

Logistic regression works by finding the best fit line that separates the data into two classes. The best fit line is determined by minimizing the residual sum of squares (RSS), which is a measure of the difference between the predicted and actual values.

Once the best fit line is determined, the model can then be used to predict the probability of an event occurring. This is done by taking the estimated probability of an event occurring and multiplying it by the probability of an event not occurring. The result is the estimated probability of an event occurring.

Logistic regression is a powerful tool for predicting and explaining the likelihood of an event occurring. It is often used in fields such as healthcare, finance, and marketing.

Use cases of logistic regression in detail.

1. Credit Risk Analysis: Logistic regression can be used to estimate the probability of default of a loan or credit card debt. By analyzing historical data of borrowers, the model can identify patterns that can be used to predict the chances of a person defaulting on a loan.
2. Predictive Maintenance: Predictive maintenance is a process of using data collected from machines to detect potential issues and optimize performance. Logistic regression can be used in predictive maintenance to identify patterns in the data that indicate a potential malfunction or failure.
3. Security Threat Detection: Logistic regression can be used to detect security threats by analyzing data from a variety of sources, such as network traffic logs, system logs, and user activity logs. The model can be used to identify malicious behavior and alert administrators.
4. Medical Diagnosis: Logistic regression can be used to diagnose medical conditions by analyzing patient data. The model can identify patterns in the data that indicate the presence of a particular illness or disease.
5. Fraud Detection: Logistic regression can be used to detect fraud by analyzing financial data. The model can identify patterns that indicate suspicious activity, such as unusual transactions or large amounts of money being transferred.

15.4 Model Evaluation

If the output is continuous variable, a measure such as the mean squared forecast error can be calculated for the test sample. For now lets assume one output and $y_i$denotes its true value for observation i, whereas $\hat{y}_i$ denotes its predicted value.

MSFE (Mean squared forecast error) = $\frac{1}{n_{test}}\sum(Y-\hat{y})^2$

Alternative forecast error aggregation is the mean absolute forecast error, where the absolute values are taken in the formula instead of the squares.

When the output is variable is a binary categorical, a common way to evaluate the model is through calculations based on a 2 X 2 confusion matrix, showing possible outcomes and whether the predicted answer was correct.

Illustration (GARP Curriculum book)

For example, suppose that we constructed a model to calculate the probability that a firm will pay a dividend in the following year or not based on a sample of 1,000 firms, of which 600 did pay and 400 did not. We would establish a threshold value of the probability, Z, which would allow the estimated probabilities to be translated into a 0–1, as discussed in the section on logistic regression. We could then set up the confusion matrix such as the following:

		PREDICTION	
		Pay dividend	Not pay dividend
OUTCOME	Pay dividend	432(43.2%) - TP	168(16.8%) – FN
	No Dividend	121 (12.1%) - FN	279(27.9%) – TN

The confusion matrix would have the same structure as long as outcome variable is binary.

Four elements of the confusion matrix as follows

1. True positive: The model predicted a positive outcome, and it was indeed positive. (TP)
2. False negative: The model predicted a negative outcome, but it was positive. (FN)
3. False positive: The model predicted a positive outcome, but it was negative. (FP)
4. True negative: The model predicted a negative outcome, and it was indeed negative. (TN)

Based on these four elements, we could specify several performance metrics,

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN} = 71.1\%$

Precision = $\frac{TP}{TP+FP} = 78.1\%$

Recall = $TP/(TP + FN) = 72\%$

Error Rate = 1 – Accuracy rate = 28.9%

There is a tradeoff between the true positive and false positive rate when setting Z that is comparable to that between type I and type II error when selecting the significance level to employ in hypothesis testing.

15.5 Decision Trees

Decision Trees are a supervised learning algorithm used for classification and regression tasks. A decision tree is a flowchart-like structure, where each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Decision Trees have several advantages over other classification algorithms, including the ability to handle both numerical and categorical data, the ability to handle missing values, and the ability to handle multi-output problems. Additionally, they can be used in areas such as medical diagnosis and credit scoring.

Decision Tree models are created by splitting the training data into subsets based on an attribute value test. The splits are chosen to maximize the information gain of each split. The information gain is a measure of the decrease in entropy, which is the measure of the amount of randomness or disorder in the system. The goal of a Decision Tree is to minimize entropy and maximize information gain.

Decision Trees are often used in conjunction with other algorithms, such as support vector machines, to improve the accuracy of the models.

CART: Classification and Regression Trees (CART) is a machine learning technique used to develop predictive models for both classification and regression problems. It is a type of

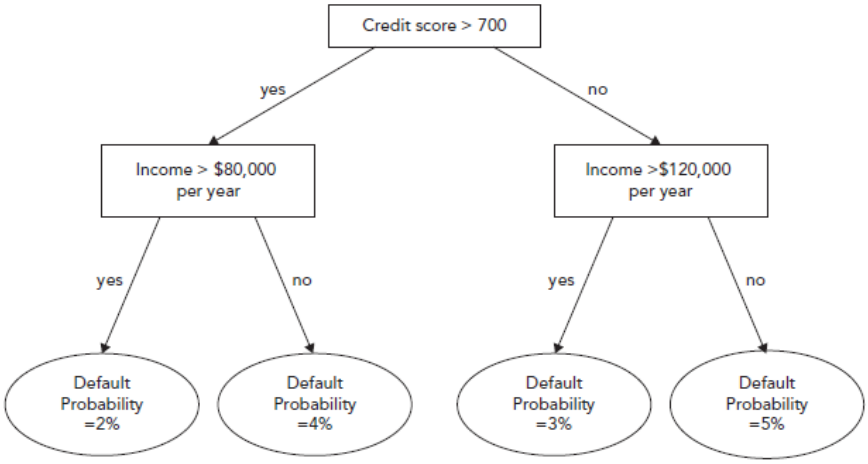
supervised learning algorithm where the goal is to construct a model that accurately categorizes a set of data points.

CART works by recursively partitioning or splitting a dataset into two or more distinct sub-datasets along predetermined features or attributes. This process is repeated until the datasets reach a point where they are homogeneous or contain only data points of the same class.

The model is constructed by choosing the feature that best divides the data into the most homogeneous sub-sets and then repeating the process for each sub-set. The split is based on a measure of impurity, such as entropy or Gini impurity, which measures how well the data points are separated in the feature.

Once the tree is complete, it can be used to make predictions on new data points by following the path down the tree that is most similar to the new data point. The final prediction is based on the values of the target variable in the data points that are reached at the end of the path.

Decision Tree



To explain how the tree is constructed, we need to introduce the concept of information gain associated with a feature. This is a measure of the extent to which uncertainty is reduced by obtaining information about the feature. The feature considered at each node is the one that maximizes the information gain. The two most widely used measures of information gain are entropy and the Gini coefficient.

Entropy is a measure of disorder and by construction, it lies between 0 and 1. The other measure is Gini. Gini and entropy are two measures of impurity used in decision tree learning. The Gini impurity measures the probability of misclassifying a randomly chosen item in a dataset if it were randomly labeled according to the class distribution in the dataset. The entropy measures the amount of information contained in a dataset, or the amount of disorder or randomness in a dataset. The entropy measure is higher when there is more disorder in a dataset, and lower when there is more order. Both measures are used to determine the best split point for a decision tree, where a low Gini or entropy indicates that the split should be done at that point.

Ensemble Technique

When building learning ensembles, a variety of models are used, and the results are combined into a single metamodel. First, by producing many of predictions and averaging them, model fit may be enhanced due to the "wisdom of crowds" and a phenomenon comparable to the law of large numbers. Secondly, the procedures are designed to prevent overfitting. The best model frequently outperforms itself when used in ensembles with weak learners. The approach is

simply explained using decision trees as an example, even though ensembles could involve combining any types of machine-learning models (including combining predictions or classifications from different classes of models, such as using both support vector machines (SVMs) and neural networks). We briefly discuss three ensemble approaches.

Bootstrap Aggregation

The process of bootstrapping from the training sample to produce multiple decision trees, as the name implies, and then aggregating the predictions or classifications from each tree to create a new prediction or classification is known as bagging. The steps below make up a simple bagging algorithm for a decision tree:

1. Take a sample from the whole training set. For instance, sample 10,000 from the training set of 100,000 observations.
2. Create a decision tree the standard way.
3. Many times, repeat steps 1 and 2, sampling with replacement to ensure that an observation made in one subsample is likewise made in another.
4. Calculate an average of the outcomes.

As a result of the replacement sampling used for the data, some observations won't show up at all. In that replication, the observations that were not chosen (referred to as out-of-bag data) will not have been utilised for estimate; nonetheless, they can be used to assess the performance of the model.

The sole difference between pasting and bagging is that sampling doesn't involve replacement (so that each datapoint can only be drawn at most once in any bootstrap replication). There would be a total of 10 sub-samples in pasting with 100,000 items in the training set and sub-samples of 10,000.

Random Forests

Random forest is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Random forests are made up of many individual decision trees. Each tree is a "weak learner" -- meaning it does only slightly better than random guessing. However, when many weak learners are combined, the result can be a powerful "strong learner".

When creating each individual decision tree, a random sample of the data (with replacement) is used for training. This randomness helps to make the model more robust as it reduces the likelihood of overfitting to the training data.

At each node in the decision tree, a random sample of the features is selected. This allows the model to make decisions based on a subset of the available features, adding variability and helping to reduce overfitting.

Finally, when the model is tested, the predictions from each of the decision trees are combined in some way (often by taking the mode for classification or the mean for regression).

Boosting

Boosting is an ensemble machine learning technique in which many models are trained together in order to produce a single, more accurate prediction. It is an iterative process in which multiple weak models are combined to form a strong model. The objective of boosting is to minimize the

errors of the weak models by giving more weight to the observations that are misclassified. The weak models are usually simple decision tree models.

In boosting, each model is trained on the same data set but with different weights assigned to each observation. The weights are initially set to 1/N where N is the total number of observations. Then, after each model is trained, the weights of the misclassified observations are increased while the weights of the correctly classified observations are decreased. This helps the next model focus on the misclassified observations and thus improving the overall accuracy.

Finally, the predictions from each of the models are combined using a weighted average. The weights for each model can be determined using a variety of techniques such as cross-validation or by optimizing an objective function such as AUC.

Boosting is a powerful technique for improving the accuracy of machine learning models and has been used to achieve state-of-the-art results in many areas.

15.6 K-Nearest Neighbors

K-nearest neighbors (KNN) is a supervised machine learning algorithm used for both classification and regression. In KNN, data is classified by a majority vote of its neighbors, with the data being assigned to the class most common amongst its k nearest neighbors. K nearest can be used for either classification or predicting the value a target variable.

KNN uses a lazy learning approach, meaning it does not use the training data to do any generalization. Instead, it stores the training data and waits until a new data point is to be classified. Then the algorithm calculates the distance between the new data point and each of the stored data points. The k-nearest neighbors are then determined based on the shortest distance. The new data point is assigned to the majority class amongst the k-nearest neighbors.

KNN can be used for both classification and regression problems. In classification, the output is a class membership (e.g. a type of fruit or a type of flower). In regression, the output is a real-valued number (e.g. the price of a house). KNN is a non-parametric algorithm, meaning it does not make any assumptions about the underlying data. This makes it useful for data sets where the distribution is not known.

Steps involved in typical KNN implementation

- Select a value of K and a distance measure (Euclidean or Manhattan)
- For each data point In the training sample, identify the K nearest neighbors in feature space to the point in feature space foe which a prediction is to be made.

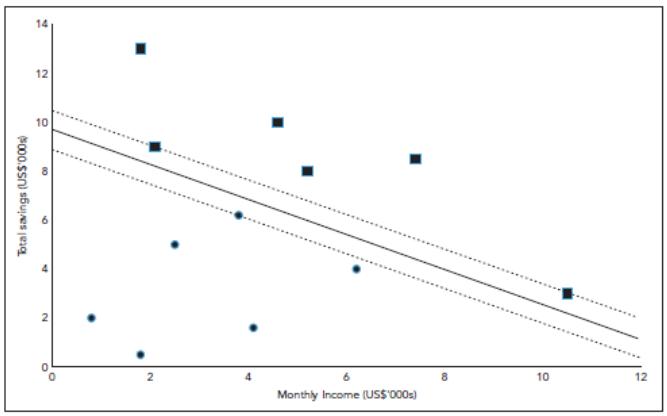
In case of classification, one might use majority voting system like forecast a class to which most of the K nearest neighbors belong. When the target value is being predicted, we can set the target equal to the average of its value for the k nearest neighbors.

The choice of K is very crucial in KNN and is based on bias variance tradeoff. If K is too large so that many neighbors are selected will result into high bias and low variance. Reverse is true for small K. Small K implies the better fit in training data but with a higher probability of overfitting. A common ***choice is to set K approximately equal to the square roof of n, the total size of the training sample.***

15.7 Support Vector Machines

Support Vector Machines (SVMs) are a type of machine-learning algorithm that can be used for both classification and regression. In classification, the aim is to find the best dividing line (or hyperplane) that separates the data into different classes. In regression, the aim is to fit the best line or curve to the data. SVMs use a technique called the kernel trick to map the data into a higher dimensional space and then build the optimal hyperplane in that space. This allows them to capture complex relationships between the data points that would otherwise be hard to detect.

To understand support vector using simple two feature example. Assume a 20-observation consisting of income of individual and saving amount in their bank. Using this information we will create hyperplane to separate the data into two groups, which will give us the information on possibility of loan default by these individuals. SVM constructs the widest path consisting of two parallel (dotted) lines , separating observations. Data points lie on the edge of the path are known as support vectors. The center line is known as separation boundary.



In this case we are using two lines but same can be extended to create a hyperplane when there are more than two features.

15.8 Neural Network

Neural Networks or Artificial Neural Networks, is a type of machine-learning model that is inspired by the structure and function of the human brain. It is composed of a large number of interconnected processing units, called neurons, which work together to perform certain tasks.

At a high level, a neural network takes in input data, processes it through a series of hidden layers, and produces an output. Each hidden layer consists of a set of neurons, and the output of one layer serves as the input for the next layer.

Each neuron in a neural network receives input from other neurons, processes this input using an activation function, and produces an output. The activation function determines how the neuron will respond to the input it receives.

The weights of the connections between neurons, as well as the biases of the individual neurons, are adjustable parameters that can be learned through training. During training, the neural network adjusts these parameters to minimize the error between the predicted output and the true output.

There are several types of neural networks, including feedforward neural networks, convolutional neural networks, recurrent neural networks, and self-organizing maps. They can be used for a wide range of tasks, including image classification, natural language processing, and time series prediction.

Key points to remember for exam.

- The most common type of ANN is a feedforward network with backpropagation, sometimes known as multi-layer perceptron. Backpropagation describes how the weights and biases are updated from iteration to another.
- The purpose of the neural network is to discover complex nonlinear relationships.

GRADIENT DESCENT ALGORITHM

In a neural network, the gradient descent algorithm is used to adjust the weights and biases of the connections between neurons in order to minimize the error between the predicted output and the true output.

To do this, the algorithm calculates the gradient of the loss function with respect to the model parameters (the weights and biases). The gradient is a vector that points in the direction of the greatest increase in the loss function. The algorithm then updates the model parameters in the opposite direction of the gradient, using the learning rate as a scaling factor.

For example, if the gradient of the loss function with respect to a particular weight is positive, this means that increasing the weight will increase the loss. The gradient descent algorithm will therefore decrease the weight in order to minimize the loss.

This process is repeated for each weight and bias in the network, until the loss function converges to a minimum. The final set of weights and biases that result from this process define the trained model.