TESTING OF HYPOTHESIS

M.Sc., STATISTICS First Year

SEMESTER-II, PAPER-II

LESSON WRITERS

Prof. G. V. S. R. Anjaneyulu
Professor of Statistics (Retd.)
Acharya Nagarjuna University

Dr. M. Vijaya Lakshmi
Associate Professor
Department of Basic Sciences
Vishnu Institute of Technology
Bhimavaram

Dr. K. Kalyani
Assistant Professor
Dept. of Mathematics & Statistics,
Vignan's Foundations For Science,
Technology & Research,
Guntur

Dr. Syed Jilani
Guest Faculty
Department of Statistics
Acharya Nagarjuna University

EDITOR

Prof. G. V. S. R. Anjaneyulu
Professor of Statistics (Retd.)
Acharya Nagarjuna University

ACADEMIC ADVISOR

Prof. G. V. S. R. Anjaneyulu
Professor of Statistics (Retd.)
Acharya Nagarjuna University

DIRECTOR, I/c.

Prof. V. Venkateswarlu
M.A., M.P.S., M.S.W., M.Phil., Ph.D.

Centre for Distance Education
Acharya Nagarjuna University
Nagarjuna Nagar 522 510

Ph: 0863-2346222, 2346208
0863- 2346259 (Study Material)
Website www.anucde.info
E-mail: anucdedirector@gmail.com

M.Sc., STATISTICS : Testing of Hypothesis

First Edition : 2025

No. of Copies :

This book is exclusively prepared for the use of students of M.Sc., STATISTICS Centre for Distance Education, Acharya Nagarjuna University and this book is meant for limited circulation only.

Published by:

Prof. V. VENKATESWARLU
Director, I/c
Centre for Distance Education,
Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A+' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the door step of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K. Gangadhara Rao
M.Tech., Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University.

M.Sc. – Statistics Syllabus

SEMESTER-II

202ST24: Testing of Hypothesis

UNIT-I:

Tests of hypotheses, concept of critical region, critical function, two kinds of errors, power function, level of significance, MP and UMP tests, Neyman Pearson lemma, Randomized and Non Randomized tests.

UNIT-II:

Generalized NP-lemma, UMP tests for simple null hypothesis against one sided alternatives, and for one sided null against one sided alternative in one parameter exponential family, extension of these results to distributions with MLR property, nonexistence of UMP test for simple null against two sided alternatives in one parameter exponential family.

UNIT-III:

UMP unbiased tests and LMP tests. Similar regions, Neyman structure, Likelihood ratio test, properties of LR test, asymptotic distribution of LR test.

UNIT-IV:

Chi-square and kolmogorov Smirnov tests for goodness of fit, Kendall's tau statistic, Kruskal- Wallis test, Friedman's two-way analysis of variance by ranks, Bartlett's test for homogeneity of variances, chi-square test for homogeneity of correlation coefficients, F-test for homogeneity of regression coefficients, variance stabilizing transformation and large sample tests.

UNIT-V:

Notion of sequential tests, SPRT, Wald's fundamental identity, relation between the quantities A,B, alpha and beta, OC and ASN functions of SPRT, application to binomial, Poisson and normal distributions, efficiency of a sequential test.

BOOKS FOR STUDY:

1) Statistical Inference by H.C, Saxena & Surendran
2) An outline of Statistical Theory vol.2 by A.M. Goon and B. Das Gupta.
3) An Introduction to probability and Mathematical Statistics by V.K. Rohatgi.
4) Mathematical Statistics- Parimal Mukopadhyay(1996), New Central Book Agency (P)Ltd., Calcutta.

BOOKS FOR REFERENCES:

1) Advanced Theory of Statistics VOL.II by M.G. Kendall & A. Stuart.
2) Introduction to Mathematical Statistics by R.V. Hogg & A.T. Craig.
3) Linear Statistical Inference and applications by C.R. Rao.

SET-I **MODEL QUESTION PAPER** **(202ST24)**

M.Sc. DEGREE EXAMINATION

SECOND SEMESTER

STATISTICS

202ST24 :: TESTING OF HYPOTHESIS

Time: 3 hours **Maximum: 70 marks**

ANSWER ONE QUESTION FROM EACH UNIT

(Each question carries equal marks)

UNIT – I

1. (a) State and prove the Neyman–Pearson fundamental lemma.
   (b) Using NP lemma, derive the most powerful test for testing H0: $\theta = \theta 0$ versus H1: $\theta = \theta 1$.

(OR)

2 (a) Define critical function and critical region with an example.
   (b) Explain Type I and Type II errors. Derive an expression for the power function for a normal mean test.

UNIT – II

3. (a) Explain the concept of Monotone Likelihood Ratio (MLR).
   (b) Show that a distribution family with MLR admits a UMP test for a one-sided alternative.

(OR)

4. (a) Define unbiased test and unbiased critical region.
   (b) Prove that an MP critical region is not necessarily unbiased.

UNIT – III

5. (a) Describe the asymptotic distribution of the likelihood ratio statistic.
   (b) Apply the LR test to test the mean of a normal population with unknown variance.

(OR)

6. (a) Define similar regions and explain their importance in testing.
   (b) Explain the concept of Neyman structures with an illustration.

UNIT – IV

7. (a) Explain Kruskal–Wallis test. Derive its test statistic and state its distribution
   (b) Explain the chi-square test for homogeneity of variances.

(OR)

8. (a) What is a non-parametric test? Describe its merits and demerits.
   (b) Explain the sign test and median test and compare them.

UNIT – V

9. (a) Explain the theoretical construction of SPRT.
   (b) Derive the approximate expressions for α and β in SPRT.

(OR)

10. (a) Describe Wald's fundamental identity.
    (b) Explain the applications of OC and ASN functions in sequential testing.

CONTENTS

LESSON -1

BASIC OF TESTING OF HYPOTHESIS

OBJECTIVES:

By the end of this lesson, students will be able to:

- Define and explain statistical hypotheses.
- Distinguish between null and alternative hypotheses.
- Describe the concept of a critical region and its role in decision-making.
- Understand the idea of a critical function and how it represents a test rule.
- Explain Type I and Type II errors with examples.
- Interpret and apply the level of significance in hypothesis testing.
- Identify the essential components of a hypothesis testing procedure.

STRUCTURE

1.1. INTRODUCTION TO HYPOTHESIS TESTING

Many times, we strongly believe that certain results or claims are true. However, when we collect a sample from the population, the observed data may not fully support our belief. This disagreement may occur due to two possibilities:

Our original belief or assumption may actually be **incorrect**, or

The sample we selected may be **unusual or one-sided**, purely by chance.

Therefore, we need **statistical tests** to distinguish between these two situations. These tests determine whether the observed difference in data can be attributed to chance variations or whether the difference is too large to be explained by randomness alone.

If the difference cannot reasonably be explained by chance, it is called **statistically significant**, and the corresponding procedures are known as **tests of significance**. The overall methodology is referred to as **Testing of Hypothesis**.

Setting up and testing hypotheses is a crucial part of **statistical inference**. Usually, a hypothesis is a claim or theory proposed either because it is believed to be true or because it serves as a basis for further investigation. However, such claims must be supported by data.

For example:

- A pharmaceutical company may claim that a new drug works better than the existing one.
- A manufacturer may claim that the average lifetime of light bulbs is 2,000 hours.
- An education researcher may claim that a new teaching method improves performance.

In all such cases, we simplify the question into two competing claims or hypotheses:

The **null hypothesis**, denoted by H0, and

The **alternative hypothesis**, denoted by H1.

These hypotheses are **not treated equally**. Special consideration is given to the null hypothesis. We reject H0 *only when the evidence against it is strong enough*.

Two common situations arise:

(i) Testing to challenge a specific claim

Often the experiment aims to disprove or reject a claim.

Example:

H0: There is **no difference** in taste between Coke and Diet Coke.

H1: There **is a difference** in taste.

The null hypothesis is rejected only if the evidence from the sample is compelling.

(ii) Testing a claim assumed to be true

Sometimes a hypothesis is treated as true unless evidence contradicts it.

Example:
A company claims that the average potency of a tablet is 250 mg. The quality control unit tests this claim.

H0: Mean potency = 250 mg

H1: Mean potency ≠ 250 mg

Again, the burden of proof lies on sample evidence.

The process involves two competing statements:

Null Hypothesis (H_0): This is the default or status quo assumption. It represents the claim being tested.

Alternative Hypothesis (H_1 or H_a): This reflects the conclusion we seek evidence for, indicating a change or difference.

Hypothesis testing provides a **structured framework** for deciding whether observed differences in data are statistically significant or likely due to chance.

This framework is critical in various fields:

In agriculture, to evaluate the effect of a new fertilizer

In medicine, to compare the effectiveness of treatments

In manufacturing, to assess product quality control

To make these decisions, we use **test statistics**, compare them to **critical values**, and determine the likelihood of committing **errors**—namely, **Type I** and **Type II** errors. The **power of a test**, defined as the ability to correctly reject a false null hypothesis, helps evaluate test efficiency.

Furthermore, the **Neyman-Pearson Lemma** provides the foundation for constructing the **most powerful tests**, especially for comparing simple hypotheses. When optimal tests are needed across a range of alternatives, we seek **uniformly most powerful (UMP)** tests.

1.2 CONCEPT OF HYPOTHESES (NULL & ALTERNATIVE)

A statistical hypothesis is an assumption or claim about a numerical characteristic of a population, known as a parameter. This might involve claims such as “the average income is ₹30,000,” “the defect rate is 5%,” or “the mean lifespan of bulbs is 800 hours.” Such hypotheses cannot be verified directly because the entire population is rarely observed, so we use sample data to test whether the claim appears to be true or not.

1.2.1 Null Hypothesis (H_0)

The **null hypothesis**, denoted by H0, is a fundamental component of hypothesis testing. It is the statement that assumes **no effect, no difference, and no change** exists in the population. In other words, it reflects the current belief, the standard condition, or the baseline situation that is considered true unless the sample provides convincing evidence to the contrary.

The null hypothesis serves as a starting point for statistical testing because it provides a specific, testable claim about a population parameter—for example, “the average performance has not improved,” “the new medicine is no more effective than the existing one,” or “there is no relationship between two variables.” By assuming H0 to be true, we calculate the likelihood of obtaining the observed sample data purely by chance.

The null hypothesis is treated with **special importance** because it is only rejected when the data present sufficiently strong evidence against it. This cautious approach prevents us from making false claims of improvement or change based on random fluctuations in the sample. In practice, we compare the observed data with what would be expected if the null hypothesis were true; if the observed outcome is extremely unlikely under H0, then we reject it in favor of the alternative hypothesis.

Thus, the null hypothesis acts as the **reference point** in every statistical test, helping us evaluate whether the sample provides enough proof to support the claim of an effect or difference.

Examples:

$H0: \mu = 50 \rightarrow$ The average lifespan is 50 years.

$H0: p = 0.3 \rightarrow$ The proportion of defective items is 30%.

$H0: \mu_1 = \mu_2 \rightarrow$ Two teaching methods produce the same average score

The null hypothesis usually includes the equality sign (=).

1.2.2 Alternative Hypothesis (H_1 or H_a)

The **alternative hypothesis**, denoted by H1 or Ha, is the statement that proposes the presence of an **effect, change, or difference** in the population. It represents the claim we want to investigate or provide evidence for. In contrast to the null hypothesis—which assumes no effect—the alternative hypothesis suggests that something meaningful or significant is occurring.

The alternative hypothesis is accepted **only when the sample data provide strong enough evidence to reject the null hypothesis**. It reflects what we expect or hope to demonstrate based on theory, prior research, or practical considerations. For example, a researcher may believe that a new teaching method improves student performance, or a company may expect that a new machine produces fewer defective items. These beliefs are expressed through the alternative hypothesis.

Depending on the nature of the research question, the alternative hypothesis may take different forms:

One-sided (one-tailed): suggests that the parameter is either greater than or less than a specific value.

Example: $H1: \mu > 50$ (mean has increased)

Two-sided (two-tailed): suggests that the parameter is simply different from a specified value, without specifying direction.

Example: $H1: \mu \neq 50$ H1 :μ□= 50

The alternative hypothesis guides the direction of the test and determines the shape of the critical region. It is the statement that the statistical test ultimately seeks to support by providing evidence strong enough to contradict the null hypothesis.

Forms of alternative hypothesis:

Left-tailed: $H1: \mu < \mu 0$

Right-tailed: $H1: \mu > \mu 0$

Two-tailed: $H1: \mu \neq \mu 0$

Real-life examples:

A pharmaceutical company wants to prove a new drug is better:

$H0: \mu_{new} = \mu_{old}$ Vs$H1: \mu_{new} > \mu_{old}$

A company claims machine lifespan is at least 5 years

$H0: \mu \geq 5$ Vs $H1: \mu < 5$

A psychologist tests whether two groups differ:

$H0: \mu 1 = \mu 2$Vs $H1: \mu 1 \neq \mu 2$

The entire process of hypothesis testing relies on comparing the results obtained from a sample with what we would typically expect to observe if the **null hypothesis** were actually true. In other words, we begin by assuming that the null hypothesis is correct and then examine whether the sample data are consistent with this assumption.

If the data closely follow the expected pattern under the null hypothesis, we do not have sufficient grounds to reject it. This suggests that any differences between the sample and the hypothesized value are likely due to normal sampling variability.

On the other hand, if the observed data deviate substantially from what the null hypothesis predicts—so much so that such a result would occur only rarely by chance—it indicates that the null hypothesis may not be a reasonable explanation. In such cases, we take the observed evidence as strong enough to reject the null hypothesis in favor of the alternative.

Thus, hypothesis testing is fundamentally about determining whether the observed differences can be explained by **random chance** or whether they point toward a **real effect or change**. This comparison between observed sample outcomes and expected outcomes under H0 forms the basis of our decision-making in statistical inference.

1.3 CRITICAL REGION

The **critical region**, also known as the **rejection region**, is one of the most important concepts in hypothesis testing. It represents the set of values of a test statistic for which we

decide to **reject the null hypothesis**. In simple terms, it is the portion of the sampling distribution where observed results

If the sample result falls in the critical region → reject H0.
If the sample result does not fall in the critical region → fail to reject H0.

Definition:

A critical region is a specific portion of the sample space that plays a central role in hypothesis testing. It is defined as the set of all sample outcomes that are considered sufficiently extreme or unusual if the null hypothesis H0 were actually true. The probability of the test statistic falling in this region

Illustrative example:

Suppose we test, H0: $\mu = 100$

using the Z-test at 5% significance.

The critical region for a two-tailed test is:

Reject H0 if $Z < -1.96$ or $Z > 1.96$

Here, the critical region includes extreme values unlikely under H0.

Real-life interpretation:

In real-world situations such as **clinical drug trials**, the concept of the critical region becomes especially meaningful. When researchers test whether a new drug is more effective than an existing treatment, they collect data on patient outcomes—such as reduction in symptoms, improvement in health scores, or recovery rates. Under the assumption that the new drug has **no real advantage** (the null hypothesis), the results should fall within a range that can reasonably be attributed to normal biological variability or random fluctuations among patients.

The **critical region** represents those sample outcomes that are so extreme—either showing **much greater improvement** or, in some cases, **far worse outcomes**—that it becomes highly unlikely they occurred merely by chance. If the observed data for the new drug fall within this critical region, the evidence suggests that the drug's effect is too large (or too small) to be explained by randomness alone.

In such situations, we reject the null hypothesis and conclude that the new drug **truly differs** in effectiveness. Thus, the critical region helps researchers make scientifically sound decisions by identifying results that indicate a genuine effect rather than routine variation.

1.4 CRITICAL FUNCTION

A **critical function** provides an alternative and more formal way of representing a hypothesis test. Instead of describing the test in terms of specific critical regions or thresholds, the

critical function expresses the test as a mathematical rule that assigns a decision—either to reject or not reject the null hypothesis—for every possible sample outcome.

It is typically written as a function $\phi(x)$, where xxx represents the observed sample or test statistic.

Instead of describing a rejection region, we define a function $\varphi(x)$:

$\phi(x)=\{1, x \in$ critical region (reject H0)

$0, x \notin$ critical region (accept H0)$\}$

$\phi(x)=\{1,0, x \in$ critical region (reject H0)

For randomized tests,

$0<\phi(x)<1$

means we reject H0 with some probability between 0 and 1.

Why critical function?

- Required for theoretical proofs (like Neyman–Pearson Lemma).
- Useful when exact significance cannot be achieved with a fixed critical region.

Small example:

Suppose the critical region is $Z > 1.645$.

Then the critical function is:

$\varphi(x) = 1$ if $Z > 1.645$

$\varphi(x) = 0$ otherwise

This gives an exact size $\alpha = 0.05$.

1.5 Types of Errors (Type I & Type II)

In a hypothesis test, a **Type-I error** occurs when the null hypothesis H0H_0H0 is rejected even though it is actually true. In other words, we conclude that there is an effect or difference when, in reality, none exists.

For example, in a clinical trial comparing a new drug with an existing one, the null hypothesis might state that the new drug is **no more effective** than the current drug. A Type-I error would occur if the analysis leads us to conclude that the two drugs differ in effectiveness when, in fact, they do not. This represents a *false positive* result.

Table 12.1 summarizes the possible outcomes of any hypothesis test:

Decision	Reject (H0)	Do Not Reject (H0)
Truth: (H0)	Type-I Error	Correct Decision
Truth: (H1)	Correct Decision	Type-II Error

A Type-I error is often considered more serious than a Type-II error, especially in fields such as medicine or public safety, because it leads us to claim a discovery or effect that is not truly present. Therefore, hypothesis testing procedures are designed to ensure that the probability of committing a Type-I error is kept very low. This probability is never exactly zero but is carefully controlled.

The probability of making a Type-I error is known as the **significance level**, denoted by:

P(Type-I Error)=α

Common values include 0.05, 0.01, or 0.10, depending on how strict the test needs to be.

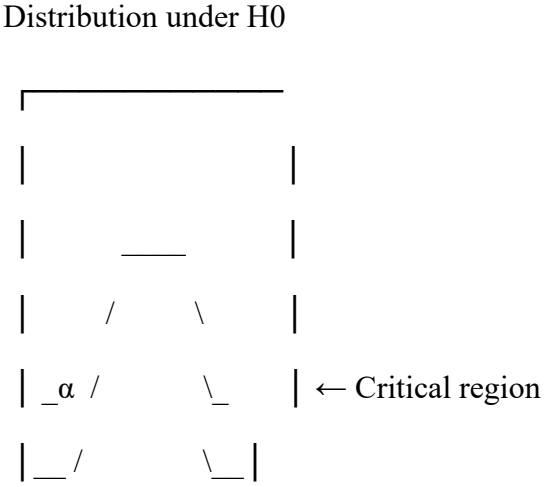
On the other hand, a **Type-II error** occurs when we fail to reject the null hypothesis even though it is false. This may happen if the sample size is too small to detect the true difference, especially when the true value lies close to the hypothesized value. The probability of a Type-II error, denoted by β\betaβ, is generally harder to calculate exactly.

Importantly, for any given test and sample size, **Type-I and Type-II errors are inversely related**. Reducing the chance of one usually increases the chance of the other. For example, if we choose a very small α (making it very hard to reject H0), we increase the likelihood of missing a real effect, leading to more Type-II errors.

A Type-I error is sometimes referred to as an **error of the first kind**, while a Type-II error is called an **error of the second kind**.

Graphical Interpretation (Bell Curve Diagram)

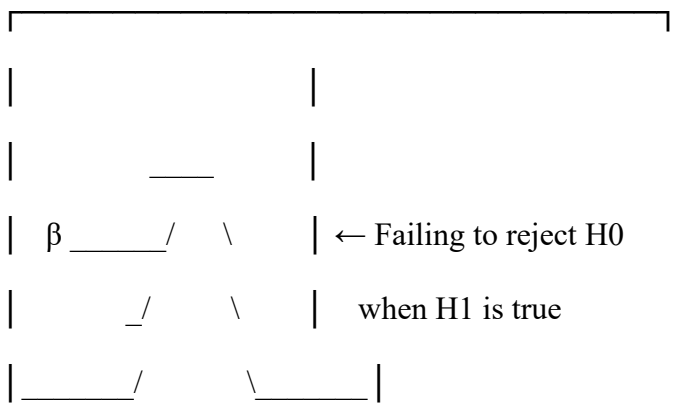
Type-I Error Area (α) Under H0



The shaded region α is the **Type-I error**: rejecting H0 when it is true.

Type-II Error Area (β) Under H1

Distribution under H1



The shaded region β is the **Type-II error**.

Relationship Between Errors

If we make α smaller (stricter test), Type-II error β often **increases**.

If we reduce β (more sensitive test), α often **increases**.

The only way to reduce **both** errors is to **increase sample size**.

This inverse relationship is crucial in designing powerful tests.

Real-life example:

Approving a medicine that does NOT work.

Convicting an innocent person.

Type II Error (False Negative)

Failing to reject H0 when H0 is false.

Denoted by β.

Real-life example:

Rejecting a life-saving drug that actually works.

Letting a guilty person go free.

Power of a Test

Power=1−β

Probability of correctly rejecting a false null hypothesis.

Important Note:

Lower α → Higher β (and lower power).

Lower β → Often requires larger sample size.

1.6 LEVEL OF SIGNIFICANCE

If we choose a significance level of **α = 0.05**, it means that we are accepting a **5% risk** of rejecting the null hypothesis even when it is actually true. In other words, we allow a small probability—5 out of 100 decisions—of making a **Type-I error**, also known as a *false positive*.

This choice reflects our willingness to tolerate a limited amount of error in order to detect meaningful effects. A 5% significance level is commonly used in many scientific fields because it strikes a balance between being too strict (which may miss real effects) and too lenient (which may lead to false claims).

Thus, setting α = 0.05 means that only sample results falling in the most unusual 5% of outcomes assuming the null hypothesis is correct—will be considered strong enough evidence to reject H0..

Selecting the significance level α\alphaα is one of the most important decisions in hypothesis testing because it determines **how strict the test will be** in deciding whether to reject the null hypothesis H0H_0H0. The value of α reflects the amount of **risk of a Type-I error** (false positive) that the researcher is willing to tolerate.

Different situations demand different levels of caution, so the choice of α depends on the **context**, **severity of error**, and **practical importance** of the decision.

1. When the Consequences Are Serious → Choose a Small α

In high-stakes situations, rejecting a true null hypothesis could lead to harmful or dangerous consequences. In these cases, we want to **minimize the chance of a false positive**, so we choose a very small α, such as **0.01, 0.005, or even 0.001**.

Examples:

(a) Medical Research

Suppose a new drug is being tested.

H0: The new drug is **no better** than the existing drug.

If we wrongly reject H0 (Type-I error), we might approve a drug that does *not* work—or worse, causes harm.

To avoid this danger, researchers choose **α = 0.01** or smaller.

(b) Aviation and Engineering Safety

When testing the safety of aircraft parts, bridges, elevators, or nuclear equipment:

A Type-I error could cause approval of a component that is defective.

Human lives depend on accuracy.

Therefore, a very small α, such as **0.001**, is used.

(c) Environmental Protection

Testing whether water, air, or food is contaminated.

False approval could expose people to toxins.

Hence, only **very strong evidence** should reject H0.

2. When the Consequences Are Less Serious → Choose a Moderate α (0.05 or 0.10)

In many everyday business, social science, or educational studies, the cost of a Type-I error is lower. Here, researchers are willing to accept a slightly higher risk of a false positive.

Common choices: **α = 0.05 or 0.10**

Examples:

(a) Market Research

A company tests whether a new advertisement is more effective.

A Type-I error may cost money, but not lives.

α = 0.05 is usually acceptable.

(b) Education Studies

A school tests whether a new teaching method improves scores.

The impact of being wrong is small (short-term experimental effect).

α = 0.10 might be reasonable.

(c) Customer Satisfaction Surveys

Testing whether customer satisfaction improved after a new policy.

A false conclusion has mild consequences.

Thus, α = 0.05 or 0.10 is used.

3. Trade-Off Between α and β
When choosing α, we must also consider **Type-II error (β)**.

If α decreases (more strict), β generally increases.

If α increases (less strict), β decreases.

So the choice should balance both errors depending on the situation.

4. Standard Scientific Practice
In many fields (biology, psychology, economics), **α = 0.05** has become a convention.

It represents a compromise between being too strict and too lenient.

It allows reasonable sensitivity while keeping false positives at an acceptable level.

However, this is not a rule—researchers adjust α depending on the **importance of the decision**.

5. Summary: Why Choose α Carefully?
α determines the strictness of your test.

Small α → more caution, fewer false positives.

Larger α → more sensitivity, fewer false negatives.

The choice depends on **risk**, **context**, **ethics**, and **practicality**.

Example:

A court uses α close to 0 → reducing the chance of convicting an innocent person.

1.7 CONCLUSION

In this lesson, we learned the basic framework that forms the foundation of statistical hypothesis testing. First, we understood that **hypotheses** are statements about unknown population parameters—such as the mean, proportion, or variance—and these statements can be tested using sample data. Hypotheses help us translate real-life questions into statistical form so they can be objectively examined.

The **null hypothesis (**H0**)** represents the position that there is *no change*, *no difference*, or *no effect*. It acts as a benchmark or reference point. The **alternative hypothesis (**H1**)**, on the other hand, represents the statement we want to investigate. It typically reflects the presence of an effect, improvement, reduction, or difference. Hypothesis testing is essentially a decision-making process between these two competing claims.

We also learned that the **critical region** plays a central role in decision-making. It consists of the values of the test statistic that are considered too unlikely if the null hypothesis were true. When the observed data fall in this region, we reject H0, concluding that the evidence suggests a real effect or difference. Complementing this, the **critical function** provides a

mathematical rule that determines whether a particular sample outcome should lead to rejection or acceptance of the null hypothesis. Together, these tools allow us to convert statistical reasoning into a clear decision rule.

Because hypothesis testing relies on sample data rather than complete population information, decisions are subject to uncertainty. As a result, two types of errors can occur:

Type I error, where a true null hypothesis is wrongly rejected, and

Type II error, where a false null hypothesis fails to be rejected.

The **significance level (α)** is a key element in hypothesis testing because it specifies the maximum probability of committing a Type I error. By choosing α (commonly 0.05 or 0.01), we control how strict our decision rule should be. A smaller α makes us more cautious in rejecting the null hypothesis.

These fundamental ideas—hypotheses, critical regions, decision functions, error types, and significance level—form the theoretical basis for more advanced topics in statistical inference. Concepts such as **Most Powerful (MP) tests**, **Uniformly Most Powerful (UMP) tests**, the **Neyman–Pearson lemma**, and **likelihood ratio tests** rely deeply on the principles introduced in this lesson. Even modern approaches such as **Bayesian decision theory** build on these ideas, though from a different philosophical viewpoint.

Thus, this lesson provides the essential groundwork for understanding how statistical decisions are made, how accuracy and reliability are measured, and how optimal tests are developed in more advanced chapters.

1.8 SELF-ASSESSMENT QUESTIONS

1. Define a statistical hypothesis with one example.
2. What is the difference between H0 and H1? Give examples.
3. Explain the meaning of the critical region in hypothesis testing.
4. What is a critical function? How is it related to the critical region?
5. Distinguish between Type I and Type II errors with real-life examples.
6. Explain the significance level α.
7. Why is it impossible to reduce both α and β at the same time without increasing sample size?
8. Create a real-life situation where a Type II error is more serious than a Type I error.

1.9 SUGGESTED READING

1. **Hogg & Tanis**, *Probability and Statistical Inference*.
2. **Gibbons & Chakraborti**, *Nonparametric Statistical Inference*.
3. **Casella & Berger**, *Statistical Inference*.
4. **Mood, Graybill & Boes**, *Introduction to the Theory of Statistics*.
5. **SP Gupta**, *Statistical Methods*.
6. **Goon, Gupta & Dasgupta**, *Fundamentals of Statistics*.

Dr. G V S R Anjaneyulu

LESSON -2

POWER FUNCTION & TEST PERFORMANCE

OBJECTIVES:

By the end of this lesson, students will be able to:

- Define Most Powerful (MP) tests and state their importance.
- Understand the concept of maximizing power under a fixed significance level.
- Explain the definition and properties of Uniformly Most Powerful (UMP) tests.
- Describe the limitations of UMP tests, especially for two-sided alternatives.
- Apply MP and UMP concepts to simple distributional problems.
- Compare MP and UMP tests and identify when each is applicable.

STRUCTURE

2.1 INTRODUCTION

In the previous lesson, we learned how hypothesis testing provides a systematic framework for making decisions about population parameters using sample data. We formulated the null and alternative hypotheses, defined critical regions, and explored the meaning of Type I and Type II errors. These concepts form the foundation of all statistical testing.

However, simply defining a hypothesis test and ensuring that the probability of committing a Type I error (α) remains within acceptable limits is **not sufficient**. A well-designed test must do more than avoid false positives. It must also be capable of **detecting incorrect null hypotheses** whenever they truly are false.

In practical terms, this means that when the null hypothesis does not hold, the test should correctly reject it with **high probability**. A test that frequently fails to identify a false null hypothesis is said to have **low power**, which can lead to misleading or inconclusive results. For example, a medical test with low power may fail to detect that a new treatment is genuinely effective, or a quality control test may fail to identify a defective production process.

Therefore, an important goal in hypothesis testing is to evaluate **how effective** a test is at identifying real differences or effects. This motivates the study of concepts such as: the **power function**, which shows how the probability of rejecting H0H_0H0 changes for different parameter values, the **size** of a test, which formalizes the maximum risk of incorrectly rejecting a true null hypothesis, **power curves**, which provide a graphical understanding of test performance, and the various **factors that influence power**, such as sample size, variance, effect size, and significance level.

By the end of this lesson, you should be able to:

Goal 1: Understand how the **power function** measures the sensitivity of a statistical test and indicates how likely the test is to detect false null hypotheses.

Goal 2: Define and clearly distinguish between the **size** of a test and the **power** of a test, and understand their roles in hypothesis testing.

Goal 3: Interpret a **power curve** and explain what it reveals about the performance and effectiveness of a statistical test.

Goal 4: Identify the major **factors that affect power**—such as sample size, effect size, variability, and significance level—and understand how each factor strengthens or weakens a test.

Goal 5: Recognize why it is important to design tests with **high power**, particularly in fields like scientific research, medicine, engineering, and quality control.

Goal 6: Apply the concepts of power, size, and power curves to **compare different statistical tests** and to plan studies where reliable detection of effects is essential.

Overall, this lesson focuses not only on controlling errors but also on ensuring that statistical tests are genuinely **effective tools for discovering truth** in uncertain situations.

To evaluate how well a test performs in this respect, statisticians study:

1. Power Function

2. Size of a Test

3. Power Curve

4. Factors that influence power

These concepts help us compare different tests and choose the most effective one. They are essential for designing studies and experiments where sensitivity and reliability are important.

For example, in clinical research, a test with low power may fail to detect that a useful drug works. Similarly, in manufacturing, a quality control test with poor power may overlook defects and allow faulty products to reach customers.

Thus, the study of power is fundamental in practical statistics.

2.2 POWER FUNCTION

2.2.1 Definition and Meaning

The **power function** shows the probability that a test will reject the null hypothesis for every possible value of the population parameter. It tells us **how sensitive** a test is to detecting real differences. A good test has **low power** near the null value (to avoid false positives) and **high power** when the true parameter moves away from the null (to detect actual effects). Thus, the power function gives a complete picture of the test's ability to correctly identify false null hypotheses.

Formally, if θ is the true parameter, the power function is:

$\pi(\theta) = P\theta(\text{Reject } H0)$

Interpretation:

For values of θ **under** H0 → the power function gives the **Type I error probability (α)** For values of $\theta \in H1$ ** → the power function gives the **probability of correctly rejecting** H0. This is the **power** of the test.

Thus, the power function evaluates the **performance** of a test across all possible true states of nature.

2.2.2 Characteristics of a Power Function

A **power function** is one that reflects how a good statistical test should perform under different parameter values. The following conditions describe its desirable behavior:

1. It should be low for all parameter values under **H0**.

When the null hypothesis H0 is true, the test should rarely reject it. This means the power function should have **low values** (close to α, the significance level) for all parameter values consistent with the null hypothesis. If the power were high under H0, it would mean the test is frequently rejecting a true null hypothesis, leading to many Type-I

errors. A stable, low power function under H0H_0H0 ensures that the test is **reliable and conservative** when the null is true.

2. It should be high for parameter values under **H1**.

When the alternative hypothesis H1 is true, the test should correctly detect the difference and reject the null hypothesis. Therefore, the power function should take **high values** (close to 1) for parameter values in the alternative region. This reflects a low probability of Type-II error. A high power under H1 indicates that the test is **effective and sensitive** in identifying real departures from the null hypothesis.

3. The function should increase as the true value moves further away from the null value.

In most practical situations, the effect becomes easier to detect when the true parameter value is far from what the null hypothesis claims. Thus, the power function should **increase smoothly** as the parameter moves away from the null value in the direction of the alternative. Near the null value, the power is close to α. As we move further away, the power rises, ultimately approaching 1. This behavior shows that the test becomes **more powerful** and **more likely to detect true differences** as the deviation from the null hypothesis becomes larger.

Summary

A well-behaved power function should:

Stay low under H0 → avoids false positives

Rise high under H1 → detects true effects

Increase as deviation from H0H_0H0 grows → reflects improved detectability

Such a power function indicates that the test is both **statistically valid** (controls errors correctly) and **practically useful** (detects differences when they exist).

2.2.3 A Simple Example (Bernoulli Model)

Let X∼Bernoulli(p).

Test:

H0:p = 0.4 Vs H1:p = 0.7

Suppose the critical region is X=1 (i.e., reject H0 if X = 1).

Then,

$\pi(p) = P(X = 1) = p$

Thus:

$\pi(0.4) = 0.4 \rightarrow$ Type I error probability

$\pi(0.7) = 0.7 \rightarrow$ Power at p=0.7

This demonstrates how power changes with parameter values.

2.2.4 Real-Life Example (Quality Control)

A manufacturer claims that the defect rate of products is only 3%. A test is designed to detect if the defect rate has increased.

$H0: p = 0.03$ Vs $H1: p > 0.03$

If the power at p=0.05 is 0.90, it means:

> "If the true defect rate is 5%, the test will detect this increase 90% of the time."

This is crucial for maintaining product quality and avoiding customer dissatisfaction.

2.3 SIZE OF A TEST

2.3.1 Definition

The **size** of a test is the maximum value of the power function **under the null hypothesis**:

$$\text{Size} = \sup_{\theta \in \Theta 0} \pi(\theta)$$

In simple vs. simple hypotheses, size = α.

2.3.2 Why Size Matters

Size tells us:

- How much risk we take of wrongly rejecting a true null hypothesis.
- Whether the test respects the significance level (α).
- How conservative or liberal the test is.
- A test with size greater than α violates the acceptable error limit.

2.3.3 Size vs. Significance Level

Although the **size** of a test and the **significance level (α)** are closely related, they are not exactly the same. Understanding the difference between them is essential for evaluating whether a test is properly designed and valid.

Significance Level (α)

The **significance level**, usually denoted by α, is the **pre-decided limit** on the probability of committing a **Type I error**—rejecting a true null hypothesis.

- It is chosen *before* conducting the test.
- It reflects how cautious we want to be.
- Common choices are 0.05, 0.01, and 0.10.

Significance level is essentially the **researcher's tolerance for risk**.

Example:
If $\alpha = 0.05$, we allow a 5% chance of wrongly rejecting H0.

Size of a Test

The **size** of a test is the **actual** maximum probability that the test will reject the null hypothesis when it is true.

Formally:

$$\text{Size} = \sup_{\theta \in \Theta 0} P\theta \text{ (Reject H0)}$$

- This depends on the **structure of the test**, not on our intended choice.
- It tells us whether the test respects the intended significance level.
- For composite hypotheses, size is the **largest Type I error probability** among all parameter values satisfying H0.

Why the Difference Matters

Even if we set $\alpha = 0.05$, the test may not actually achieve this value. For example:

The true maximum Type I error could be **less than** 0.05 (making the test conservative).

Or it could be **more than** 0.05 (meaning the test violates the allowed error rate).

Thus, the size is the **realized Type I error probability**, while α is the **intended limit**.

Validity Condition

For a statistical test to be considered **valid**, the size must not exceed the chosen significance level:

Size $\leq \alpha$

This ensures that the test does not reject the null hypothesis too often when it is actually true.

Example to Clarify the Difference

Suppose the significance level is set at $\alpha = 0.05$. However, because of the way the critical region is constructed (for example, due to discreteness of data), the actual probability of rejecting H0 might be:

For some parameter values under H0: 0.03

For others: 0.04

For the worst case: **0.047**

Thus:

Size = $0.047 \leq 0.05$

The test is valid.

But if the maximum probability under H0H_0H0 turned out to be:

Size = $0.058 > 0.05$

Then the test **violates α** and is invalid.

Summary

Concept	**Meaning**	**Chosen vs. Actual**	**Purpose**
Significance Level (α)	Intended risk of Type I error	Chosen **before** testing	Controls strictness
Size	Actual maximum Type I error	Determined **by the test**	Validates correctness

A test is valid only if **Size ≤ α**, ensuring that the actual risk of falsely rejecting H0H_0H0 does not exceed the allowed limit.

2.3.4 Example

Let a test reject H0 if sample mean $\bar{x} > 10$ for H0:$\mu=8$.

Compute:

$\alpha = P8(\bar{x} > 10)$

This value is the size of the test.

If this exceeds the allowed α, the test must be redesigned.

2.4 POWER CURVE AND INTERPRETATION

2.4.1 Definition

A **power curve** is a visual representation of the **power function** of a statistical test. It is created by plotting the power function $\pi(\theta)$\pi(\theta)$\pi(\theta)$ on the vertical axis against different possible values of the parameter θ on the horizontal axis.

In simple terms:

Power Curve =Graph of $\pi(\theta)$ versus θ

The power curve displays how the probability of rejecting the null hypothesis changes as the true parameter value varies. It provides an immediate picture of:

- How well the test performs near the null hypothesis (where power should be low)
- How quickly the power increases as the parameter moves into the alternative region
- Whether the test is capable of detecting even small deviations from the null
- How the test compares to other candidate tests

Because graphs are easier to interpret than formulas, the power curve is a powerful tool for researchers and students to understand the **effectiveness**, **sensitivity**, and **reliability** of a statistical test.

2.4.2 Importance

It visually shows:

- How well the test can detect departures from H0
- For which values of the parameter the test is most effective
- Where the test is weak

2.4.3 Interpretation of the Curve

A **power curve** not only displays the mathematical behavior of the power function but also provides valuable insights into how effective a statistical test is in practice. Interpreting the

power curve helps us understand the strengths and weaknesses of a test across different situations.

Typically:

Near H0: power is low (close to α)

Far from H0: power increases (approaches 1)

A **steep power curve** indicates a highly sensitive test.

Here are the key points of interpretation:

1. Power Near the Null Hypothesis (**H0**)

For values of the parameter that fall under the null hypothesis, the power curve should be **low**, i.e., close to the significance level α.

- This indicates that the test is **not too aggressive** in rejecting a true null hypothesis.
- If the power is high near H0, the test is unreliable because it produces many **false positives** (Type-I errors).

Thus, a low power near H0 reflects **good Type-I error control**.

2. Power in the Alternative Region (**H1**)

When the parameter lies in the alternative region, the power curve should rise **significantly**, ideally approaching 1.

This means:

- The test becomes **more capable of correctly rejecting** a false null hypothesis.
- A power close to 1 indicates a very effective and sensitive test.

If the power remains low in the alternative region, the test has **poor ability to detect real differences**.

3. Rate at Which the Curve Rises

How fast the power curve rises as the parameter moves away from the null value tells us how **sensitive** the test is.

- A **steep rise** in the curve means the test quickly gains power and is effective at detecting even small departures from H0.

- A **flat curve** means the test is weak and may fail to detect meaningful differences, even when they exist.

This steepness is often used to compare different tests.

4. Maximum Power

- The highest value of the power curve is ideally **1** (or very close to 1). This means the test almost always rejects the null hypothesis when it is false.
- If the power never approaches 1, the test may not be suitable for practical use.

5. Comparison of Tests Using Power Curves

Power curves allow comparison of two or more tests. The test whose power curve lies **above the others** for most parameter values is usually better.

Higher curve = better performance

Lower curve = weaker test

This method is widely used in test selection and experimental design.

6. Symmetry or Direction of the Curve

For **two-sided tests**, power curves often have symmetric shapes around the null value.

For **one-sided tests**, the curve rises only in one direction.

This visual cue helps identify the type of hypothesis being tested.

Overall Interpretation

A power curve helps answer the question:
"How good is the test at detecting departures from the null hypothesis?" A good power curve should:

- Stay low near H0
- Rise steadily and steeply as the parameter moves into H1
- Approach 1 for large deviations

Clearly show the advantage of one test over another.

Thus, the power curve is a practical tool for evaluating test performance and guiding researchers in choosing the most effective statistical method.

2.4.4 Example (Normal Distribution)

To better understand how a power curve behaves, consider a test based on the **normal distribution**, where the goal is to compare the mean of a population to a known value. Suppose we are testing:

$H0:\mu = \mu0$ vs. $H1:\mu > \mu0$

Here, the test is designed to detect whether the true mean μ is **greater** than the hypothesized mean $\mu0$.

Because of the properties of the normal distribution, the test statistic (typically the sample mean or a standardized z-statistic) shifts to the right as the true mean μ increases. As a result:

1. Power increases quickly as μ moves above μ_0

When the actual mean is only slightly larger than $\mu0$\mu_0$\mu0$, the test has a moderate ability to detect this difference. But as μ becomes further and further above $\mu0$\mu_0$\mu0$, the test statistic is more likely to fall into the rejection region.

This causes the **power function** to increase. In graphical form, the power curve begins close to the significance level α at $\mu=\mu0$, then rises as μ increases.

2. A Steeply rising power curve indicates a good test

If the power curve rises **very sharply** as μ increases above μ_0, this means the test is:

- **Highly sensitive**
- Able to detect even small differences
- **Effective** at distinguishing false null hypotheses

Such a test is desirable because it has **low Type-II error** when the true mean deviates from $\mu0$.

Example:
A test with large sample size (n = 100 or more) often produces a steep power curve, because the sample mean becomes more precise.

3. A Slowly Rising Power Curve Indicates Poor Sensitivity

If the power curve rises **slowly**, the test struggles to detect differences even when μ is noticeably larger than μ_0.

This indicates:

- Low sensitivity
- High Type-II error (β)
- Weak performance

Reasons for slow rise may include:

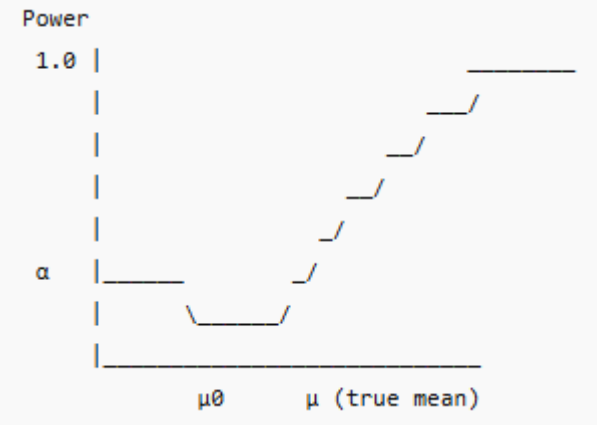
- Small sample size
- High variance
- Poorly designed test

Using a two-tailed test when a one-tailed test is appropriate

Such a test may fail to detect meaningful effects in practical applications.

4. Visual Summary

A typical power curve for a normal-distribution mean test looks like this:



The curve starts near α at $\mu = \mu_0$

- Rises as μ increases
- A steep rise = good test
- A flat rise = weak test

Conclusion of the Example

This example illustrates how the **shape** of the power curve directly reflects the **effectiveness** of the test. For normal-based tests of the mean, the curve provides a clear visual indicator of whether the test is properly detecting changes in the true mean.

2.5 FACTORS AFFECTING POWER

The power of a statistical test reflects its ability to **detect a false null hypothesis**. A test with high power correctly identifies real effects or differences more frequently. Several factors

influence the power of a test, and understanding them helps researchers design stronger, more reliable studies.

Below are the major factors that affect the power of a test:

2.5.1 Sample Size (n)

How it affects power:

The sample size is one of the most important determinants of power. As sample size increases:

- The standard error decreases
- The sampling distribution becomes narrower
- The test statistic becomes more precise
- The chance of detecting small differences increases

Result:

Larger sample size → higher power

Example:

A clinical trial with **30 patients** may fail to detect a moderate improvement caused by a new drug. The same trial with **300 patients** is far more likely to detect the true effect.

2.5.2 Effect Size

How it affects power:

Effect size is the magnitude of the true difference or the strength of the effect between the null and alternative hypothesis.

Large effect size → test easily detects the difference

Small effect size → test struggles to detect the difference

Result:

Larger effect size → higher power

Example:

Detecting a change in average weight from **60 kg to 80 kg** is easier (large effect). Detecting a change from **60 kg to 62 kg** is much harder (small effect), requiring larger sample size.

2.5.3 Population Variance (σ^2)

How it affects power:

Variance measures how spread out the data are. High variability makes it harder to detect differences, because:

- The signal (effect) is drowned by noise (random variation)
- The test statistic becomes less precise

Result:

Lower variance → higher power

Higher variance → lower power

Example:

A blood pressure study with very variable readings ($\sigma = 20$) has lower power than one with stable readings ($\sigma = 5$).

2.5.4 Significance Level (α)

How it affects power:

The significance level determines the size of the critical region. Increasing α makes it easier to reject H0.

Higher α → larger critical region → higher power

Lower α → smaller critical region → lower power

Trade-off:

- High α increases Type-I error risk
- Low α increases Type-II error risk

Result:

Increasing α increases power but increases false positives

Example:

A test with $\alpha = 0.10$ is more likely to detect an effect (higher power) than a test with $\alpha = 0.01$.

2.5.5 Type of Test (One-Tailed vs. Two-Tailed)

How it affects power:

A **one-tailed test** concentrates the entire significance level in one tail, giving:

- A larger critical region
- Higher chance of rejecting H0H_0H0 in the specified direction

A **two-tailed test** splits α into two tails, reducing power.

Result:

One-tailed test → higher power (if direction is correct)

Two-tailed test → lower power

Example:

If we know a new fertilizer can only **increase** crop yield, a one-tailed test is more powerful than a two-tailed test.

2.5.6 Shape of the Sampling Distribution

How it affects power:

Normal, symmetric, and well-behaved sampling distributions typically yield higher power because:

- The location shift under H1is easier to detect
- Critical values are well-defined

Non-normal or skewed distributions often require larger sample sizes to achieve the same power.

Result:

More regular sampling distribution → higher power

2.5.7 Measurement Precision and Study Design

How it affects power:

Improved measurement tools and better study design can reduce variability and improve accuracy.

- High-quality instruments → lower noise
- Better experimental protocols → fewer errors

Result:

Better quality data → higher power

Example:

A precise thermometer (accurate to 0.01°C) increases power compared to a rough instrument (accurate to only 1°C).

Factor	**Increases Power**	**Decreases Power**
Sample Size (n)	Larger n	Smaller n
Effect Size	Larger effect	Smaller effect
Variance (σ^2)	Low variance	High variance
Significance Level (α)	Larger α	Smaller α
Test Type	One-tailed	Two-tailed
Sampling Distribution	Normal/low noise	Skewed/high noise
Measurement Quality	High precision	Low precision

2.5.8 Summary of Factors Affecting Power

Overall Insight

A well-designed test aims for:

- High power (to detect real differences)
- Controlled α (to avoid false positives)
- Minimal β (to avoid missing real effects)

Understanding how different factors affect power allows researchers to plan studies that are **statistically strong and scientifically reliable**.

2.6 CONCLUSION

In this lesson, we studied key concepts that evaluate the effectiveness of a statistical test:

The **power function** tells us how likely the test is to reject H0H_0H0 across different parameter values. The **size** of the test measures the maximum Type I error probability. The **power curve** provides a visual understanding of test performance. Power depends on sample size, effect size, variance, significance level, and test direction. These ideas help researchers design better experiments and choose optimal statistical tests. Understanding power also prevents incorrect conclusions caused by weak tests.

2.7 SELF-ASSESSMENT QUESTIONS

1. Define the power function.
2. What is the difference between size and significance level?
3. Why is sample size important for power?
4. Explain effect size with an example.
5. Draw and interpret a power curve.
6. Explain the relationship between α and β.
7. Why do one-tailed tests have more power than two-tailed tests?
8. A test has low power. What does this imply?
9. List three factors that can increase the power of a test.
10. Calculate the power function for a Bernoulli distribution with critical region $X = 1$.

2.8 FURTHER READINGS

1. **Casella & Berger**, *Statistical Inference*
2. **Hogg & Tanis**, *Probability and Statistical Inference*
3. **Mood, Graybill & Boes**, *Introduction to the Theory of Statistics*
4. **Goon, Gupta & Dasgupta**, *Fundamentals of Statistics*
5. **Wackerly, Mendenhall & Scheaffer**, *Mathematical Statistics with Applications*
6. **Lehmann & Romano**, *Testing Statistical Hypotheses*
7. **Gibbons & Chakraborti**, *Nonparametric Statistical Inference*

Dr. G V S R Anjaneyulu

LESSON -3

MP and UMP Tests

OBJECTIVES:

By the end of this lesson, students will be able to:

- Define Most Powerful (MP) tests and state their importance.
- Understand the concept of maximizing power under a fixed significance level.
- Explain the definition and properties of Uniformly Most Powerful (UMP) tests.
- Describe the limitations of UMP tests, especially for two-sided alternatives.
- Apply MP and UMP concepts to simple distributional problems.
- Compare MP and UMP tests and identify when each is applicable.

STRUCTURE

3.1 INTRODUCTION

In hypothesis testing, there may be many different statistical tests available for examining the same hypothesis. However, these tests are **not equally effective**. Some tests are better at identifying false null hypotheses, while others may miss important differences. This brings us to the concept of **optimality** in hypothesis testing—the idea that among all possible tests, some perform better than others according to specific criteria.

One of the most important criteria used to judge the quality of a test is its **power**, which measures the probability that the test will correctly reject the null hypothesis when it is false. A test with higher power is more sensitive and more reliable in detecting true effects. The search for tests that maximize power leads to two central concepts:

- **Most Powerful (MP) Tests:** These tests provide the highest power for a specific alternative hypothesis when both the null and alternative are simple (i.e., completely specified).
- **Uniformly Most Powerful (UMP) Tests:** These tests extend the idea of MP tests to situations where the alternative is composite. A UMP test has the **highest power for all parameter values** in the alternative hypothesis, not just one specific value.

Understanding MP and UMP tests is crucial because they offer a systematic way to identify or construct the **best** possible test under given conditions. They form the theoretical backbone of classical hypothesis testing and provide a framework for comparing different statistical procedures.

By the end of this lesson, you should be able to:

Goal 1: Understand the concept of **Most Powerful (MP) tests**, including how they maximize power for simple alternatives.

Goal 2: Explain the **Neyman–Pearson Lemma**, which provides the foundation for identifying MP tests.

Goal 3: Understand the idea of **Uniformly Most

3.2 MOST POWERFUL (MP) TESTS

In hypothesis testing, we often need to decide which statistical test is *best* for detecting a false null hypothesis. The concept of **Most Powerful (MP) tests** helps us formalize this idea. MP tests are designed to **maximize the probability of correctly rejecting** H0 for a given significance level when the alternative hypothesis is true.

3.2.1 Definition of MP Tests

A statistical test is said to be **Most Powerful (MP)** for testing a simple null hypothesis against a simple alternative hypothesis if:

- It has significance level α,
- **No other test with the same significance level has higher power** at the specified alternative value.

Formally, a test with critical region C is MP of size α if:

$$P\theta 1(X \in C) \geq P\theta 1(X \in C')$$

for any other test with critical region C′satisfying $P\theta 0(X \in C') \leq \alpha$

"Among all tests that control Type I error" means that when comparing different statistical tests, we only consider those that keep the probability of rejecting a true null hypothesis at or below the pre-chosen significance level α. In other words, we compare only tests that are **valid**—tests that do *not* exceed the allowed chance of making a Type I error.

Once this condition is met, we look for the test that provides the **highest probability of rejecting** H0 when the alternative hypothesis is true. That test is called the **Most Powerful (MP)** test.

- Among all tests that keep the Type I error probability at most α, an MP test is the one that maximizes the probability of detecting a false null hypothesis.
- This ensures the test is both **statistically valid** (controls α) and **optimally sensitive** (highest power).

3.3 UNIFORMLY MOST POWERFUL (UMP) TESTS

In hypothesis testing, we often need to decide which statistical test is *best* for detecting a false null hypothesis. The concept of **Most Powerful (MP) tests** helps us formalize this idea. MP tests are designed to **maximize the probability of correctly rejecting** H0 for a given significance level when the alternative hypothesis is true.

3.4 MP Vs. UMP – COMPARISON & LIMITATIONS

This section presents a detailed comparison between Most Powerful (MP) tests and Uniformly Most Powerful (UMP) tests, outlining their conceptual foundations, practical differences, and the challenges involved in constructing them. The discussion begins by explaining how MP tests and UMP tests differ in scope: MP tests are designed for simple hypothesis scenarios, targeting maximum power at a *specific* alternative, whereas UMP tests aim to be optimal *uniformly* over an entire class of alternatives. Because of this narrower focus, MP tests are generally easier to derive, typically using the Neyman–Pearson Lemma in simple vs. simple hypothesis settings.

In contrast, UMP tests are far more difficult to obtain. They often do not exist, especially when dealing with composite or two-sided alternatives, because no single test can dominate all others across the entire range of parameter values. The section explains how this non-existence issue arises from conflicting power requirements and the absence of monotone likelihood ratios in many models. It further notes that in two-sided testing problems, constructing a UMP test is usually impossible, and one must instead rely on UMP unbiased (UMPU) tests, which balance power and fairness by ensuring that the test does not systematically favor any direction of the alternative.

Finally, the section discusses practical limitations when applying MP or UMP concepts in real-world data analysis. These include model misspecification, the presence of nuisance parameters, small-sample complications, and deviations from the theoretical assumptions that underpin optimality results. Together, these points underscore the theoretical elegance but practical constraints of MP and UMP testing frameworks.

3.5 ILLUSTRATIVE EXAMPLES

This section presents a set of carefully worked-out examples designed to reinforce the theoretical concepts of MP and UMP tests. Each example demonstrates how likelihood ratios, test statistics, and rejection regions are constructed under different probability models, highlighting the practical application of the Neyman–Pearson Lemma and the conditions under which UMP tests exist. The examples cover both continuous and discrete distributions, offering a balanced and intuitive understanding of hypothesis-testing procedures.

The first example illustrates a **Most Powerful (MP) test for the mean of a normal distribution under simple versus simple hypotheses**. Using the Neyman–Pearson Lemma, the likelihood ratio is derived explicitly, followed by the determination of the rejection region based on the observed sample mean. This example emphasizes how MP tests exploit the complete specification of both hypotheses to achieve maximum power at a particular alternative.

Next, the section provides a **Uniformly Most Powerful (UMP) test for a one-sided hypothesis about the mean of a normal distribution with known variance**. Here, the monotone likelihood ratio property allows the construction of a UMP test valid for all parameter values under the one-sided alternative. The example clarifies why UMP tests exist in this setting and how they lead to familiar z-test procedures.

To broaden applicability, the section also includes **examples involving binomial and Poisson distributions**, demonstrating how MP and UMP tests are derived in discrete settings. These examples highlight the role of probability mass functions, monotone likelihood ratios, and critical values based on cumulative distribution functions. They also show how to manage discrete rejection regions when exact significance levels are not possible.

Each example is accompanied by an **interpretation of the corresponding rejection region through likelihood ratios**—explaining intuitively what it means for observed data to "favor" the alternative hypothesis.

Where appropriate, the section incorporates **graphical illustrations**, such as plots of likelihood functions, rejection regions, and power curves. These visual aids help develop intuition about how evidence accumulates against the null hypothesis and how test optimality is reflected in the likelihood ratio structure.

Example 1: MP Test for Normal Mean (Simple vs. Simple)

Model: $X1,X2,\ldots,Xn \sim N(\mu,\sigma^2)$

Hypotheses:

$H0:\mu = \mu_0$ vs.$H1:\mu = \mu_1 (\mu_1 > \mu_0)$

Step 1: Likelihood Ratio

$$\Lambda(x)= \frac{L(\mu 0)}{L(\mu 1)} = exp\left(\frac{n}{\sigma^2}\ (\mu_1 - \mu_0)(2\bar{x} - ((\mu_1 + \mu_0)\ )\right)$$

Reject H0 when $\Lambda(x)$ is small → equivalently when $\bar{x}$ is large.

MP Rejection Region

$$\bar{x} > c$$

For size α:

$$P\mu_0(\bar{x} > c) = \alpha \Rightarrow c = \mu_0 + z_{1-\alpha}\sigma_n$$

Interpretation

The data rejects H0 when the sample mean is sufficiently larger than μ_0. This test is MP because NP Lemma applies to simple vs. simple hypotheses.

Example 2: UMP One-Sided Test for Normal Mean (Composite Alternative)

Model: Xi $\sim N(\mu,\sigma^2)$; σ known.

Hypotheses:

$H0:\mu \leq \mu 0$ vs. $H1:\mu > \mu 0$

Reason UMP Exists

Normal family has a **monotone likelihood ratio (MLR)** in $\bar{x}$.

UMP Test

Reject H0 when

$\bar{x} > \mu_0 + z_{1-\alpha}\sigma_n$

This is the standard **one-sided z-test**.

Interpretation

This test is optimal **for all** $\mu>\mu_0$, not just a specific value.

Example 3: Binomial Distribution – MP and UMP Test

Model: X~Binomial(n,p)

Case A: MP Test (Simple vs. Simple)
H0:p = p0 vs H1:p = p1, p1 > p0

Likelihood ratio:

$$\frac{P(X=x|p1)}{P(X=x|p0)} = \left(\frac{(P_1(1-P_0)}{(P_0(1-P_1)}\right)^x$$

Increasing in x.

Thus MP test:

Reject H0 if $X \geq k$

Choose k such that:

$P_{p0}(X \geq k) \leq \alpha$

Case B: UMP Test for Composite Alternative

$H0: p \leq p0$ vs $H1: p > p0$

Binomial family has MLR in X.

Thus **UMP test exists**:

Reject when X is large.

This parallels the normal example but in discrete form.

Example 4: Poisson Distribution – MP/UMP Test

Model: $X \sim Poisson(\lambda)$

Hypotheses

$H0: \lambda \leq \lambda 0$ vs $H1: \lambda > \lambda 0$

Likelihood ratio:

$$\frac{P(X=x|\lambda 1)}{P(X=x|\lambda 0)} = \left(\frac{\lambda 1}{\lambda 0}\right)^{x} \exp(-(\lambda 1-\lambda 0))$$

Increasing in x.

→ Poisson family has MLR in X.

UMP Test

Reject H0 if $X \geq k$

where k is chosen so that

$P_{\lambda 0}(X \geq k) \leq \alpha$

Example 5: Likelihood-Ratio Interpretation (Graphical Insight)

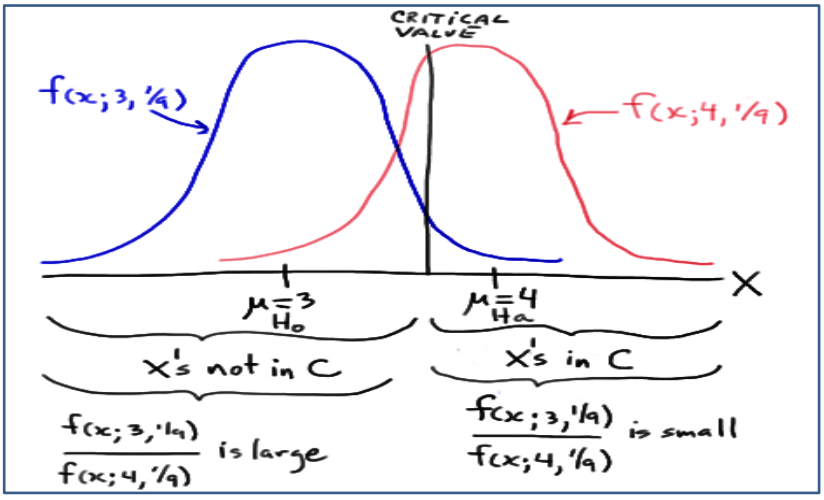
Given a sample statistic T(x), the LR test defines the rejection region:

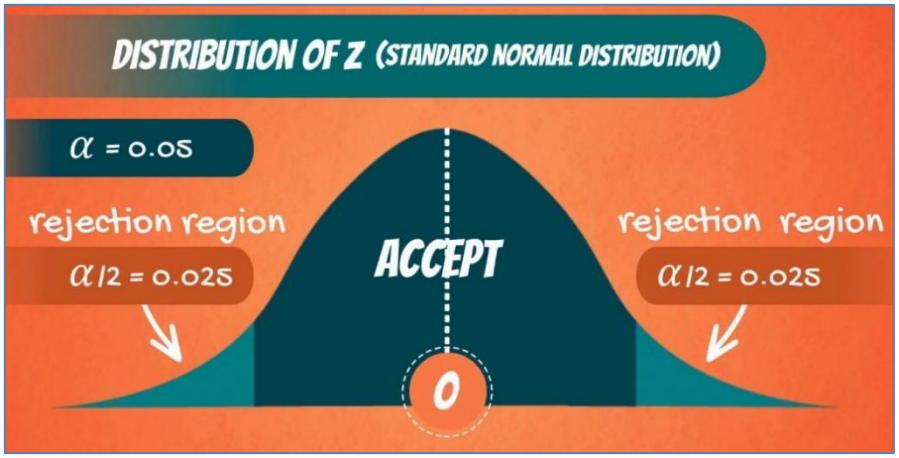
$$\frac{L(\theta 0)}{L(\theta 1)} \leq c$$

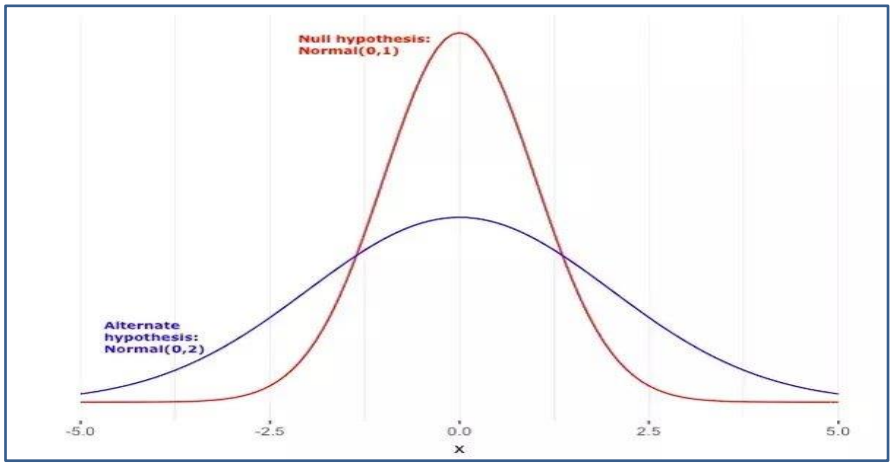
Graphically:

- Plot likelihood under H0 and H1.
- Regions where L(H1) > L(H0) indicate stronger evidence against H0.
- The rejection region corresponds to the tail where the likelihood ratio is smallest.

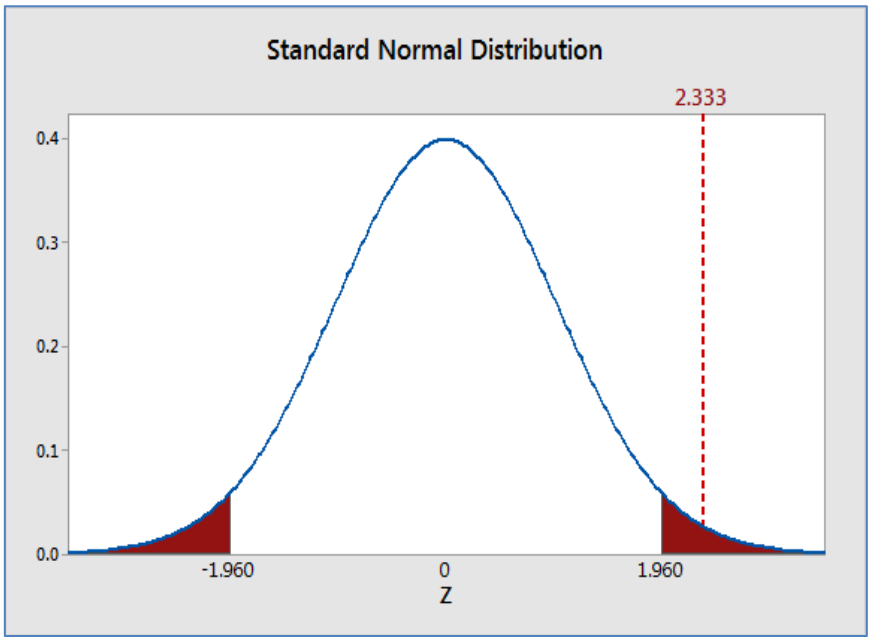
1. Likelihood Ratio – Normal Distribution (MP Test)

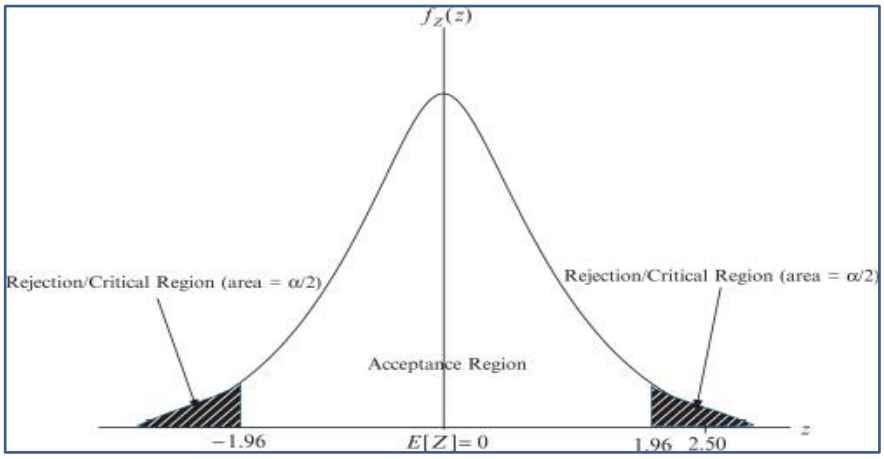


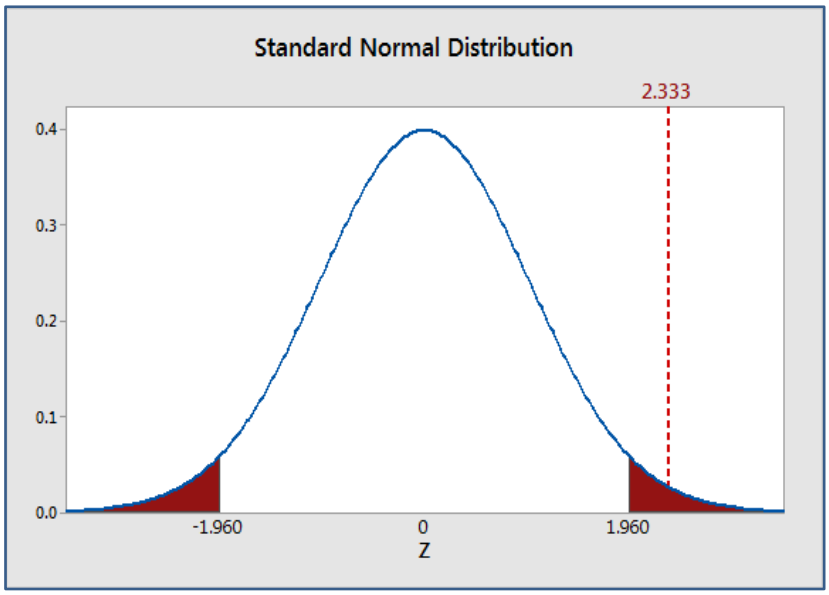




2. UMP One-Sided Test – Normal Mean (Z-Test)





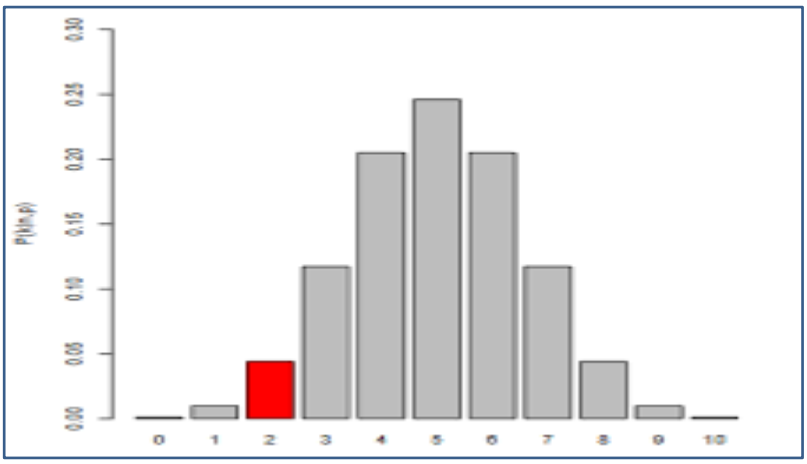


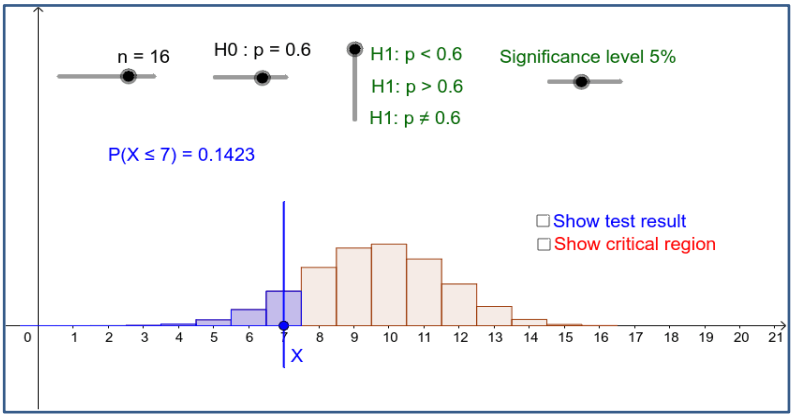
Interpretation:

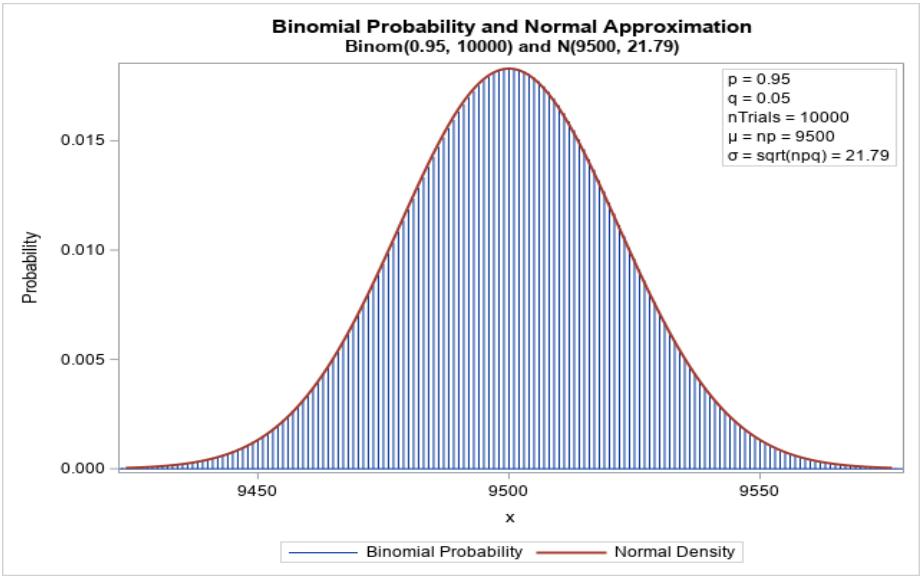
The rejection region is the right tail beyond $z_{1-\alpha}$.

This test is UMP for all $\mu > \mu_0$ due to monotone likelihood ratio.

3. Binomial MP / UMP Test – Rejection Region



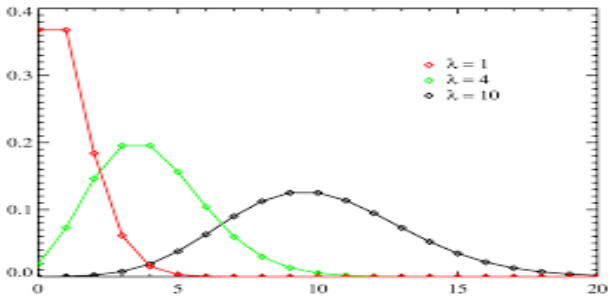




Interpretation:

- Bars on the right tail form the rejection region when p1 > p0.
- For composite alternative p>p0, the same structure becomes UMP.

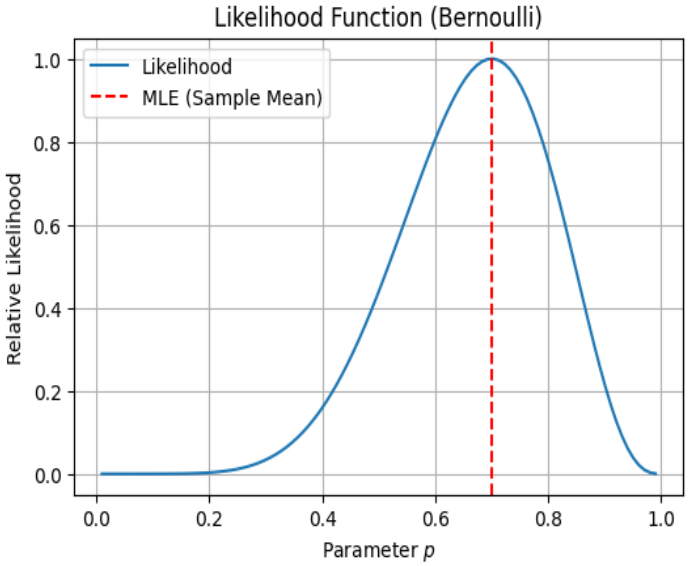
4. Poisson Distribution – UMP Test



Interpretation:

- For H1:λ > λ0, large values of X favor the alternative.
- Right-tail rejection region is used.

5. Likelihood Ratio Function – Conceptual Diagram

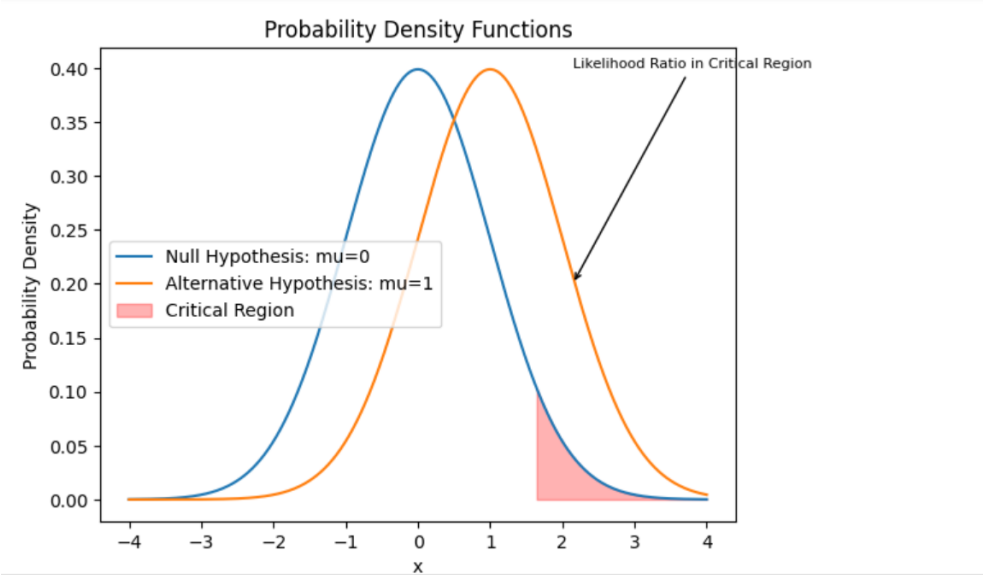


Interpretation:

- The LR test rejects when the ratio L(H0)/L(H1)L(H_0)/L(H_1)L(H0)/L(H1) is small.
- Graph shows where likelihood under the alternative surpasses the null.

MP vs. UMP Tests – Visual Summary

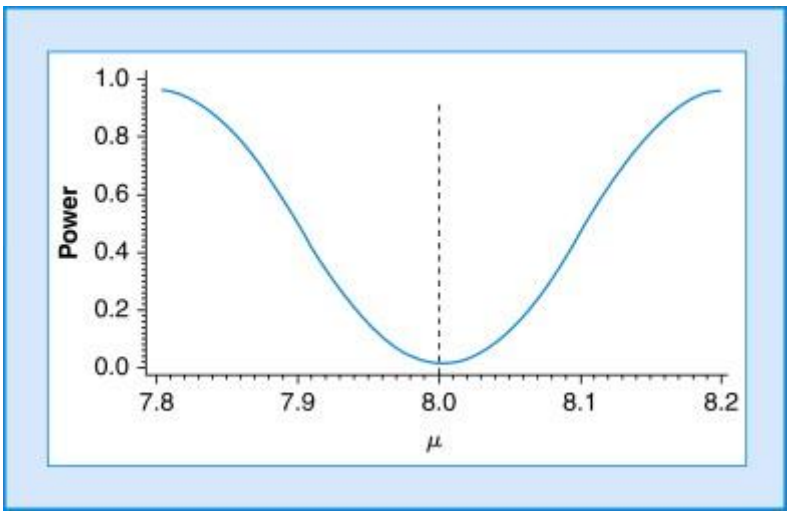
1. MP Test – Simple vs. Simple (Narrow Focus)



Key Idea

- MP tests maximize power **for one specific alternative** (e.g., μ1).
- Rejection region chosen using the likelihood ratio between H0 and H1.
- Optimal **only** at one point.

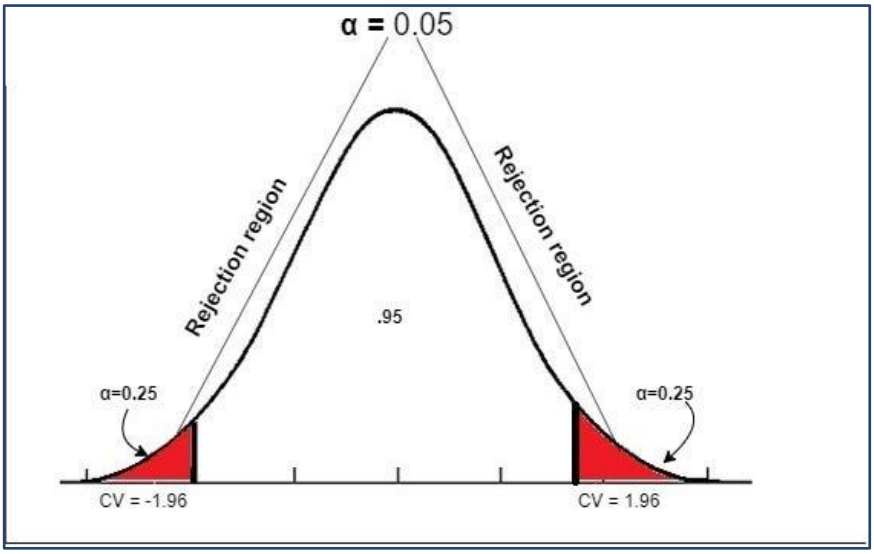
2. UMP Test – Composite Alternative (Uniform Optimality)



Key Idea

- UMP tests maximize power **uniformly for all alternatives** in a direction (e.g., all μ > μ0).
- Exists only when the family has **monotone likelihood ratio (MLR)**.
- One-sided normal mean test is the classic example.

3. Why UMP Often Does Not Exist



Key Idea

- In two-sided testing, no single test dominates everywhere.
- Power trade-offs across alternatives prevent uniform optimality.
- Often need **UMP unbiased (UMPU)** tests instead.

Summary Table

Feature	MP Test	UMP Test
Hypotheses	Simple vs simple	Simple vs composite
Optimality	At one specific parameter	Uniform across parameter range
Existence	Always (NP Lemma)	Rare (requires MLR)
Two-sided tests	Works	Usually impossible
Unbiasedness needed?	No	Often required
Interpretation	Best test for a specific alternative	Best test for all alternatives in a direction

3.6 CONCLUSION

This section brings together the central ideas developed throughout the chapter on Most Powerful (MP) and Uniformly Most Powerful (UMP) tests. MP tests are grounded in the Neyman–Pearson framework and provide the highest possible power when comparing simple hypotheses—where both the null and alternative specify a single parameter value. Their strength lies in the ease of construction and the guarantee of optimality for that specific alternative.

In contrast, UMP tests aim for a more demanding objective: maximizing power **uniformly** over an entire range of alternatives rather than at a single point. Because of this requirement, UMP tests exist only under special structural conditions in the underlying probability model—most notably the presence of a **monotone likelihood ratio (MLR)**, which ensures that the rejection region can be chosen consistently across all alternatives in the specified direction.

However, the chapter also emphasizes that UMP tests are rare in practice. They often fail to exist, especially in two-sided testing problems or in models with nuisance parameters, where no single test can dominate all others for every alternative value. As a result, real-world statistical analysis frequently involves compromises, such as relying on UMP unbiased (UMPU) tests, approximate methods, or tests chosen for robustness rather than strict optimality.

Overall, the chapter highlights both the theoretical elegance and the practical limitations of MP and UMP tests, providing a foundation for understanding more advanced topics in optimal hypothesis testing.

3.7 SELF-ASSESSMENT QUESTIONS

1. Define a Most Powerful (MP) test.
2. State the Neyman–Pearson Lemma and explain its importance in hypothesis testing.
3. What conditions are necessary for a Uniformly Most Powerful (UMP) test to exist?
4. Explain the difference between MP and UMP tests using appropriate examples.
5. Why do UMP tests often fail to exist for two-sided alternatives?
6. Explain the concept of a monotone likelihood ratio (MLR).
7. Why does MLR guarantee the existence of UMP tests in one-sided testing problems?
8. What is meant by a “simple” versus a “composite” hypothesis?
9. Give an example of an MP test in the normal distribution for simple vs. simple hypotheses.
10. Construct a UMP test for testing $H0:\mu \leq \mu 0$ vs. $H1:\mu > \mu 0$ in a normal model with known variance.
11. Explain how the rejection region is determined in an MP test using the likelihood ratio.
12. Describe how hypothesis tests change when dealing with discrete distributions such as the Binomial or Poisson.
13. Why do discrete distributions sometimes make it impossible to achieve the exact desired significance level?
14. What is a UMP unbiased (UMPU) test?
15. Why is unbiasedness required in many two-sided testing scenarios?
16. Explain why UMP tests fail in the presence of nuisance parameters.
17. What role does the Karlin–Rubin theorem play in UMP testing?
18. Give an example where no UMP test exists, and explain why.
19. Discuss the limitations of MP tests when the alternative hypothesis is composite.
20. Explain why MP tests are relatively easy to derive compared with UMP tests.
21. Explain the concept of monotone likelihood ratio.

3.8 SUGGESTED READING BOOKS

Recommended references include:

1. **Casella & Berger**, *Statistical Inference*
2. **Lehmann & Romano**, *Testing Statistical Hypotheses*
3. **Mood, Graybill & Boes**, *Introduction to the Theory of Statistics*
4. **Hogg & Tanis**, *Probability and Statistical Inference*
5. **Wackerly, Mendenhall & Scheaffer**, *Mathematical Statistics with Applications*
6. **Goon, Gupta & Dasgupta**, *Fundamentals of Statistics*
7. **Bickel & Doksum**, *Mathematical Statistics*

Dr. G V S R Anjaneyulu

LESSON -4

NEYMAN–PEARSON LEMMA & RANDOMIZED TESTS

OBJECTIVES:

By the end of this lesson, students will be able to:

- State the Neyman–Pearson Lemma and explain its theoretical significance.
- Construct MP tests for simple hypotheses using the NP Lemma.
- Describe the need for randomized tests in certain testing situations.
- Distinguish between randomized and non-randomized tests with examples.
- Use likelihood ratio forms to derive optimal tests.
- Apply NP-based test construction to real statistical problems.

STRUCTURE

4.1 INTRODUCTION

Hypothesis testing is one of the fundamental components of statistical inference, enabling researchers to make informed decisions about population characteristics based on sample data. In many situations, more than one statistical test may be available for assessing the same hypothesis, and these tests can differ considerably in terms of their effectiveness. A natural question, therefore, is: *Which test should we choose to obtain the strongest possible evidence against the null hypothesis when it is false?* This question leads us to the study of **Most Powerful (MP)** and **Uniformly Most Powerful (UMP)** tests.

MP and UMP tests belong to a class of optimal procedures that are designed to maximize the ability of a test to detect deviations from the null hypothesis. An MP test is the most effective test for discriminating between **two simple hypotheses**, but its optimality is restricted to a

single alternative parameter value. UMP tests extend this idea by seeking optimality across **an entire range of alternatives**, making them extremely desirable when they exist. However, such tests require special mathematical conditions and may not be available in many practical scenarios.

The starting point for developing MP tests is the **Neyman–Pearson Lemma**, a foundational result in statistical theory. This lemma provides a precise rule for constructing the most powerful test of a given size when testing simple hypotheses. It introduces the concept of the **likelihood ratio**, a central idea that forms the backbone of many optimal testing procedures. Because of its clarity and generality, the Neyman–Pearson framework is widely regarded as one of the most elegant and impactful contributions to modern statistical methodology.

While MP tests are straightforward to derive for simple hypotheses, real-world problems frequently involve composite hypotheses, where the parameter space includes multiple possible values. In such cases, the construction of optimal tests becomes more complicated. Some situations allow the development of UMP tests through properties such as the **monotone likelihood ratio (MLR)**, but these cases are exceptions rather than the rule. When UMP tests do not exist, statisticians must rely on alternative approaches, such as UMP unbiased tests or likelihood ratio–based methods.

This chapter also distinguishes between **randomized** and **non-randomized** tests. Randomized tests may appear theoretical, but they play an important role in situations involving discrete distributions—such as the binomial or Poisson models—where achieving an exact significance level is not always possible. In contrast, non-randomized tests are more intuitive and are typically used in continuous models like the normal distribution.

To strengthen conceptual understanding, the chapter includes detailed examples illustrating how MP tests are developed in the context of the normal, binomial, and Poisson distributions. These examples highlight the mechanics of computing likelihood ratios, identifying rejection regions, and interpreting the optimality of the resulting tests.

The chapter concludes with a set of self-assessment questions aimed at reinforcing key ideas, along with a list of suggested readings for students who wish to explore optimal testing theory more deeply.

4.2 THE NEYMAN–PEARSON LEMMA

The **Neyman–Pearson Lemma (NP Lemma)** is one of the foundational results in the theory of hypothesis testing. It provides a rigorous method for identifying the **Most Powerful (MP)** test when comparing two *simple* hypotheses. A simple hypothesis is one that specifies the parameter completely, leaving no uncertainty about its value. The hypotheses considered in the NP Lemma take the form:

$H0:\theta=\theta0$ Vs $H1:\theta=\theta1$

Statement (Informal)

Among all tests having a fixed significance level α, the test that is most powerful for distinguishing H0 from H1 is the one that rejects H0 for small values of the likelihood ratio:

$$\Lambda(x)=\frac{f(x|\theta 0)}{f(x|\theta 1)}.$$

Implications

- The test based on the likelihood ratio maximizes power for the specified alternative.
- The lemma applies only to **simple vs. simple** hypotheses.
- It gives a constructive method for identifying the rejection region.

This likelihood ratio compares how plausible the observed data x is under the null hypothesis relative to the alternative. Smaller values of $\Lambda(x)$ indicate stronger evidence against H0.

Thus, the MP test rejects H0 when:

$$\Lambda(x) \leq k,$$

where the constant k is chosen so that the test has significance level α.

Why the Likelihood Ratio?

The likelihood ratio is a natural and intuitive measure of evidence: If the data are **much more likely** under H1 than under H0, the ratio becomes small, suggesting rejection of H0. If the data are **more compatible** with H0H_0H0, the ratio remains large, and we retain the null. The ratio therefore quantifies the *relative plausibility* of the two hypotheses. **Implications of the Neyman–Pearson Lemma (Expanded)**

1. The likelihood ratio test is most powerful

The NP Lemma guarantees that, among all tests with the same level, no other test has greater power at $\theta 1$, and no alternative rejection rule performs better for this simple alternative.

This gives the likelihood ratio test a **unique optimality** that no other method can surpass for this setting.

2. Valid only for simple vs. simple hypotheses

The lemma **cannot** be applied directly when: the null hypothesis is composite (e.g., H0: $\theta \leq \theta 0$), or the alternative is composite (e.g., H1:$\theta > \theta 0$). In these cases, finding UMP tests becomes far more challenging, and sometimes impossible.

3. Provides a constructive method for deriving the rejection region

The lemma doesn't just assert that an optimal test exists—it tells us *exactly how to build it*. The steps are:

- Compute the likelihoods under H0 and H1.
- Form the likelihood ratio.
- Determine how the ratio behaves (increasing or decreasing in a statistic like $\bar{x}$.
- Use this to construct the rejection region in terms of a sample statistic.

- Choose the constant k so that the test has size α.
- This makes MP test construction highly systematic.

4. The likelihood ratio reflects strength of evidence

The NP Lemma establishes that evidence is best captured not by: the sample mean alone, or variance, or a test statistic chosen arbitrarily, but by the **likelihood ratio**, which directly compares how well the two hypotheses explain the data.

This interpretation forms the basis of: likelihood ratio tests (LRT), generalized likelihood ratio tests (GLRT), and many optimality principles in modern statistics.

Graphical Interpretation (Conceptual)

When plotting the density under H0 and H1: The rejection region corresponds to the area where the curve for H1 dominates the curve under H0. The point where the two curves intersect often marks the boundary between acceptance and rejection. The likelihood ratio is small in the region where the alternative is more plausible. This visual perspective helps understand why the NP-based rejection region is optimal.

Summary

The Neyman–Pearson Lemma: identifies the optimal test for simple hypotheses, establishes the likelihood ratio as the most informative statistic, provides a concrete recipe for constructing MP tests, and lays the foundation for the broader family of likelihood-based tests.

Neyman–Pearson Lemma — statement and proof

Let X be a random variable (or vector) with densities (or probability mass functions) f0(x) under H0 and f1(x) under H1, defined on a common measurable space (X,A). A (possibly randomized) test is a measurable function $\varphi: X \to [0,1]$ where $\varphi(x)$ is the probability of rejecting H0 at observation x.

For a test φ define its size and power as

$\alpha(\varphi) = E0[\varphi] = \int \varphi(x)\, f0(x)\, dx,$

$\beta(\varphi) = E1[\varphi] = \int \varphi(x)\, f1(x)\, dx,$

where Ei denotes expectation under fi.

We fix a significance level $0 < \alpha < 1$. The goal is to find, among all tests with size $\leq \alpha$, a test that maximizes the power $\beta(\varphi)$.

Theorem (Neyman–Pearson)

Define the likelihood ratio

$\lambda(x) = \frac{f(x/\theta_1)}{f(x/\theta_0)}$ (interpreted appropriately where f0(x)=0)

For any constant c > 0 and γ∈[0,1], consider the test φ given by

$$\varphi(x) = \begin{cases} 1, & \lambda(x)>c \\ \gamma, & \lambda(x)=c \\ 0, & \lambda(x)<c \end{cases} \qquad (1)$$

Choose c and γ so that α(φ∗) = α (this is always possible by varying c and γ; if exact equality is not possible without randomization, choose γ∈(0,1)\gamma\in(0,1)γ∈(0,1) appropriately on the boundary). Then: φ has size α.

For any test φ with α(φ) ≤ α, we have β(φ) ≤ β(φ∗). That is, φ∗ is **most powerful** among all tests of size α.

Moreover, if $P_0(\lambda(X) = c) = 0$ (no mass on the boundary), the test is nonrandomized (γ is irrelevant) and is unique (up to a.s. equivalence) among size-α MP tests.

Proof

1. Existence / construction and size

Define the sets

Ac = { x: λ(x) > c},Bc ={x: λ(x) = c},Cc ={x:λ(x) < c}.

For a given c, choose γ∈[0,1] so that the test φ∗defined by (1) satisfies

$\alpha(\varphi*) = \int_{Ac} 1 \cdot f_0(x)\,dx + \int_{Bc} \gamma \cdot f_0(x)\,dx = \alpha.$

Because $\int_{Ac} f_0$ is a decreasing right-continuous function of c and $\int Ac \cup Bc\, f_0$ is its left-limit, by varying c and then γ we can achieve any target level in the interval [0,1]. Hence we can choose c and γ that make α(φ∗) = α. (This is the place where randomization on the boundary Bc can be required if $\int_{Ac} f_0$ jumps over α)

Thus φ∗ is a valid test of size α.

2. Optimality (main inequality)

Let φ be any other test with α(φ) ≤ α. Consider the difference in powers:

$\beta(\varphi*) - \beta(\varphi) = \int(\varphi*(x)-\varphi(x))\, f_1(x)\,dx$

Fix the constant c used in φ. Multiply and subtract c times the corresponding difference in sizes:

$\beta(\varphi*)-\beta(\varphi)-c(\alpha(\varphi*)-\alpha(\varphi)) = \int(\varphi*(x)-\varphi(x))(f_1(x)-cf_0(x))\,dx. \qquad (2)$

But $\alpha(\varphi*)=\alpha$ and $\alpha(\varphi) \leq \alpha$, so $\alpha(\varphi*)-\alpha(\varphi) \geq 0$ Therefore

$$\beta(\varphi*) - \beta(\varphi) \geq \int(\varphi*(x)-\varphi(x))(f_1(x)-cf_0(x))\, dx. \quad (3)$$

Now analyze the integrand pointwise. Note that

$$f_1(x)-cf_0(x) = f_0(x)(\lambda(x)-c)$$

Consider three regions:

If $\lambda(x)>c$ then $\varphi*(x) =1$, so

$$(\varphi*(x) - \varphi(x))(f_1(x)-cf_0(x)) = (1-\varphi(x))\, f0(x)\, (\lambda(x)-c) \geq 0.$$

If $\lambda(x) < c$ then $\varphi*(x) = 0$ and $\lambda(x)-c < 0$, so

$$(\varphi*(x) - \varphi(x))(f_1(x) - cf_0(x)) = -\varphi(x)\, f_0(x)\, (\lambda(x) - c) \geq 0.$$

If $\lambda(x)=c$ then $f_1(x)-cf_0(x) = 0$

Thus the integrand $(\varphi*-\varphi)(f_1-cf_0)$ is **everywhere nonnegative**, and hence the integral is ≥ 0. Combining with (3) we obtain

$$\beta(\varphi*)-\beta(\varphi) \geq 0.$$

So $\beta(\varphi*) \geq \beta(\varphi)$. Because φ was any test with $\alpha(\varphi) \leq \alpha$, $\varphi*$ is most powerful at level α.

This proves optimality

3. Comments on equality and uniqueness

If $P_0(\lambda(X) = c) = 0$ (i.e., the boundary set Bc has zero f_0 -mass), we can take γ equal to 0 or 1 without changing $\alpha(\varphi*)$. In that case $\varphi*$ is nonrandomized and the inequality above is strict wherever φ differs from $\varphi*$ on a set of positive f_0-mass, which implies uniqueness (modulo null sets).

If $P_0(\lambda(X)=c) > 0$, different choices of γ (or different assignments on the boundary) can produce different MP tests, all having the same size and the same power. Randomization is therefore essential in discrete situations to achieve exact size and to make the lemma cover all possibilities.

Remarks and intuition

The key trick is (2): subtracting c times the size-difference converts the difference in powers into an integral where the integer and has a definite sign because of how $\varphi*$ was built in terms of where f1exceeds cf_0. This is the heart of the NP argument.

The likelihood ratio $\lambda(x) = \lambda(x) = \frac{f1(x)}{f0(x)}$ orders observations by how much they favor H1 over H0. The NP test simply rejects when the evidence (as measured by λ) is large — or equivalently when $\Lambda(x) = \frac{f1}{f0}$ is small.

Randomization appears only to exactly hit a prescribed size α when the distribution under H0 has point-masses (discrete models); for continuous models the boundary usually has probability zero and no randomization is needed.

Example (discrete, showing need for randomization)

Suppose X takes integer values and under H0 we have $P_0(X \geq 5) = 0.06$, $P_0(X \geq 4) = 0.12$. If α=0.1, no nonrandomized cut of the form {reject if $X \geq k$} attains α exactly (because 0.06 and 0.12 sandwich 0.10). The NP solution randomly rejects when X=4 with appropriate probability $\gamma \in (0,1)$ chosen so that $0.06 + \gamma P0(X=4) = 0.10$

Conclusion

The Neyman–Pearson Lemma provides a complete and practical framework for constructing the **most powerful test** when comparing two simple hypotheses. It shows that the key quantity for making an optimal decision is the **likelihood ratio**, which evaluates how much more likely the observed data are under the alternative hypothesis than under the null. According to the lemma, the best strategy is to **reject the null hypothesis whenever the likelihood ratio $\frac{f1(x)}{f0(x)}$ becomes sufficiently large**, or equivalently, whenever the inverse ratio $\frac{f0(x)}{f1(x)}$ becomes sufficiently small. In other words, we reject the null precisely in those regions of the sample space where the evidence most strongly favors the alternative.

What makes the lemma especially powerful is that it offers not just an abstract optimality result but a **constructive recipe**:

- Compute the likelihood ratio.
- Identify the values of the data for which this ratio is smallest under H_0.
- Choose the rejection region by selecting a threshold that ensures the test has the correct significance level.

This guarantees that no other test of the same size can achieve higher power against the specified alternative.

In situations involving **discrete distributions**, the exact significance level may not be attainable using a simple cut-off rule. The NP Lemma resolves this issue by allowing **randomization at the boundary**, meaning that for certain observations, the test rejects the null hypothesis with a probability between 0 and 1. This ensures that the test's size is controlled exactly at the prescribed level α. Randomization is therefore not a complication of the theory but a necessary component for achieving precise optimality when the probability distribution has jumps.

The mathematical proof of the lemma is both elegant and accessible. It relies on analysing the **sign of a key integrand** that represents the difference in power between any candidate test

and the NP test. By showing that this integrand is always nonnegative, the proof demonstrates that the NP test dominates all competitors. This argument does not require complex mathematical tools—only careful algebra and an understanding of how the likelihood ratio orders the data.

Overall, the Neyman–Pearson Lemma stands as one of the most influential results in statistics. It not only establishes a gold standard for optimal testing but also lays the groundwork for the broader class of likelihood-based methods, including generalized likelihood ratio tests and many modern developments in statistical decision theory.

4.3 CONSTRUCTING MP TESTS USING THE NP LEMMA

To build an MP test, the following steps are typically used:

- **Write down the likelihood function** under H0 and H1.
- **Form the likelihood ratio:** $\Lambda(x) = \frac{L(\theta 0)}{L(\theta 1)}$
- **Determine how the ratio changes** with the sample statistic.
- **Define the rejection rule**: reject H0 when $\Lambda(x) \leq k$.
- **Find the constant** k so that the test has size α.

4.4 RANDOMIZED TESTS

In some cases—especially with discrete distributions—it is impossible to select a non-randomized test that achieves an exact significance level α.
To handle this, **randomized tests** assign probability values between 0 and 1 to the decision at specific boundary points.

When randomization is used

- Hypothesis tests for **binomial or Poisson** distributions.
- Cases where the significance level cannot be achieved exactly using cut-off values.

Illustrative Example (Binomial)

If $X \sim Bin(n,p_0)$ and $Pp0\ (X \geq k) < \alpha < Pp0\ (X \geq k-1)$, then the test may randomize at $X = k-1$:

Reject H_0 with probability γ, where γ is chosen to make the total size equal to α. Randomized tests are mathematically essential, although rarely used in day-to-day applied work.

4.5 NON-RANDOMIZED TESTS

A **non-randomized test** always makes a definitive decision—either reject or fail to reject the null.

These tests are:

- Preferred in practice,
- Perfectly suitable for continuous distributions,

- Easy to interpret and implement.

General form

$\phi(x)=\{1, x\in R(\text{reject})$

$0, x\notin R(\text{fail to reject})$

In continuous models (normal, exponential), non-randomized tests can exactly satisfy the significance level without any randomization.

4.6 APPLICATIONS AND EXAMPLES

Examples Using Likelihood Ratio

Simple Example (Bernoulli Distribution)

Let X~Bernoulli(θ) test:

H_0: θ = 0.5

H_1: θ = 0.7

Sample: one observation x = 0 or 1

Step 1: Write pmf (probability mass function):

$f(x;\theta)=\theta^x(1-\theta)^{1-x}$

Step 2: Likelihood ratio:

$$\lambda((x) = \frac{f(x;0.7)}{f(x;0.5)} = \begin{cases} \frac{0.7}{0.5} = 1.4, \text{ if } x = 1 \\ \frac{0.3}{0.5} = 0.6, \text{if } x = 0 \end{cases}$$

Step 3: Apply the rule:

If λ((x)>k, reject H_0.

Choose k so that total probability of rejecting H_0 when H_0 is true equals α.

Example 1: Binomial Distribution (Discrete Case)

Problem:

Let X~Binomial(n=10,θ).
We want to test:

H_0: θ = 0.5

H_1: θ = 0.7

At significance level α = 0.05

Step 1: Likelihood Function

$f(x;\theta)=\binom{10}{x}\theta^x(1-\theta)^{10-x}$

Under H_0: $\theta = 0.5$
Under H_1: $\theta = 0.7$

Step 2: Likelihood Ratio

$\lambda((x) = \frac{f(x;0.7)}{f(x;0.5)} = \left(\frac{0.7}{0.5}\right)^x \left(\frac{0.3}{0.5}\right)^{10-x}$

This is an increasing function in **x** ⇒ Larger x means more evidence **against H_0**

Step 3: Critical Region

Use critical region: **Reject H_0 if $X \geq k$**

Find value of **k** such that:

$P(X \geq k \sim H0:\theta=0.5) \leq 0.05$

Using Binomial(10, 0.5) table:

$P(X \geq 8)=P(8)+P(9)+P(10)=0.044+0.010+0.001=0.055$

$P(X \geq 9)=0.010+0.001=0.011$

So, set **k = 9** to ensure $\alpha \approx 0.011 < 0.05$

Conclusion:

Reject H_0 if $X \geq 9$

This is the **most powerful test** of size ≤ 0.05

Example 2: Normal Distribution (Continuous Case)

Problem:

Let $X \sim N(\mu, \sigma^2=1)$.
We want to test:

H_0: $\mu = 0$

H_1: $\mu = 1$

Based on a **single observation x**, at $\alpha = 0.05$

Step 1: Likelihood Function

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

Under H_0: $\mu = 0$
Under H_1: $\mu = 1$

Step 2: Likelihood Ratio

$$\lambda(x) = \frac{f(x;1)}{f(x;0)} = e^{\frac{-1}{2}(x-1)^2-(x)^2}$$

This is an **increasing function in x**
$\Rightarrow$ Larger x gives more support to H_1

Step 3: Critical Region

Choose k such that:

$\Lambda(x) = \exp(x-0.5) > c$

$\Rightarrow x > \ln(c) + 0.5$

We need to find **critical value x_0** such that:

$P(X > x_0 \sim H_0: \mu = 0) = 0.05$

From standard normal table:

$Z_{0.05}=1.645$

Conclusion:
Reject H_0 if $x > 1.645$

This is the **most powerful test** of size $\alpha = 0.05$ for testing $\mu = 0$ vs $\mu = 1$.

4.7 CONCLUSION

This chapter developed the theoretical foundation for most powerful tests, showing how the Neyman–Pearson Lemma provides a clear and optimal way to handle simple hypotheses. The chapter differentiated between randomized and non-randomized tests, explained why randomization appears in discrete models, and showed how MP tests are constructed using the likelihood ratio.

Although MP tests are the most efficient for simple hypotheses, they have limited use when dealing with composite alternatives. The existence of UMP tests requires special mathematical structure—such as the presence of a monotone likelihood ratio—so they are not available for many commonly encountered testing problems. Consequently, practical applications often rely on approximate procedures, unbiased tests, or likelihood ratio–based methods.

4.8 SELF-ASSESSMENT QUESTIONS

1. Define a Most Powerful (MP) test.
2. State the Neyman–Pearson Lemma in your own words.
3. What is meant by a simple hypothesis?

4. Describe the idea behind the likelihood ratio.
5. When does a MP test become a UMP test?
6. Explain why UMP tests are rare for two-sided alternatives.
7. What is a randomized test? Give an example.
8. Why are non-randomized tests sufficient for continuous distributions?
9. Construct the MP test for testing H0:μ=μ0 vs. H1:μ=μ1 .
10. What is the monotone likelihood ratio property?
11. Why is MLR important for the existence of UMP tests?
12. Compare the rejection regions for MP and UMP tests.

4.9 SUGGESTED READING – BRIEF GUIDE

1. Lehmann & Romano – ***Testing Statistical Hypotheses***
2. Casella & Berger – ***Statistical Inference***
3. Hogg, Tanis & Zimmerman – ***Probability and Statistical Inference***
4. Mood, Graybill & Boes – ***Theory of Statistics***
5. Wasserman – ***All of Statistics***

Dr. M.Vijaya Lakshmi

LESSON -5

GENERALIZED NEYMAN–PEARSON LEMMA

OBJECTIVES:

By the end of Lesson 5, students will be able to:

- Understand the limitations of the classical NP Lemma for composite hypotheses.
- Define and explain the **Generalized Neyman–Pearson (GNP) Lemma**.
- Distinguish between simple and composite null/alternative hypotheses.
- Apply the GNP Lemma to derive most powerful tests in constrained parameter spaces.
- Explain the power function in composite settings using supremum/infimum arguments.
- Identify situations where GNP Lemma is useful in theoretical test construction.

STRUCTURE

5.1 INTRODUCTION TO COMPOSITE HYPOTHESES

In statistical hypothesis testing, the form of the hypotheses plays a crucial role in determining how a test is constructed and how optimality is defined. A hypothesis is called **simple** if it specifies the parameter of interest **exactly**, leaving no ambiguity about the distribution under that hypothesis. For example,

$$H0:\theta = \theta 0$$

is a simple hypothesis because the parameter takes one specific value.

However, in real-world applications, hypotheses rarely specify a single value. Instead, they represent a **range** or **set** of possible parameter values. Such hypotheses are known as **composite hypotheses**.

Examples of Composite Hypotheses

Consider the hypotheses:

$H0:\theta \leq \theta 0$ vs $H1:\theta > \theta 0$.

These are composite because:

- Under H0, the parameter θ can take **any value less than or equal to** $\theta 0$
- Under H1, the parameter can be **any value greater than** $\theta 0$.

Thus, each hypothesis corresponds to a **set of parameter values**, not a single value.

Why Composite Hypotheses Are More Complex

Composite hypotheses are more difficult to handle because a test must perform well **uniformly** over many possible parameter values. For simple hypotheses, optimality is straightforward: compare the likelihoods under two fixed parameter values. But for composite hypotheses, several challenges arise:

Multiple Likelihood Functions

Under a composite null or alternative, the likelihood function changes depending on the specific value of the parameter. This means the likelihood ratio is no longer a single function but a family of functions.

Type I Error must be controlled for all values

For a test to maintain a given significance level α, it must satisfy:

$$\sup_{\theta \in \Theta 0} P_\theta(\text{Reject H0}) \leq \alpha$$

That is, the worst-case Type I error across all parameter values in the null space must be controlled.

Power Must Be Good for Every Alternative

Power is no longer checked at one point but across an entire set:

$$\inf_{\theta \in \Theta 1} P_\theta(\text{Reject H0})$$

This means the test must perform uniformly well across all alternatives.

Optimal Tests may not exist

Because of conflicting performance requirements over a range of parameter values, there may not exist a test that is uniformly superior across all alternatives.

Relation to the Generalized Neyman–Pearson Lemma

The classical Neyman–Pearson Lemma is powerful but applies only when *both* hypotheses are simple. It identifies a test that maximizes power at a **single alternative parameter value**.

Composite hypotheses require an extension of this framework.
This leads to the **Generalized Neyman–Pearson (GNP) Lemma**, which provides conditions under which a test can be considered uniformly most powerful over sets of parameter values.

The GNP Lemma shifts the optimality conditions from **pointwise comparisons** to **uniform comparisons** across entire parameter ranges. This is essential for constructing UMP tests in one-sided testing problems, particularly within structured families such as the exponential family.

5.2 REVIEW OF CLASSICAL NEYMAN–PEARSON LEMMA (EXPANDED EXPLANATION)

The classical **Neyman–Pearson Lemma** provides the most powerful test for distinguishing between two *simple* hypotheses:

$H0:\theta = \theta0$ $H1:\theta = \theta1$

The key result can be summarized as:

Among all level-α tests, the likelihood ratio test (LRT) that rejects H0 when

$\Lambda(x) = \frac{f(x|\theta0)}{f(x|\theta1)} \leq k$ is the **Most Powerful (MP)** test.

What the Classical NP Lemma Guarantees

- **Optimality** at **one** specific alternative value ($\theta1$)
- **Exact control** of Type I error
- A **specific rejection region** determined by the likelihood ratio

What the NP Lemma Cannot Do

It **cannot** tell us how to construct tests when the null or alternative involves **multiple** parameter values.

It **cannot** guarantee uniform power across a range of alternatives.

It **cannot** determine UMP tests for composite hypotheses.

Thus, although the NP Lemma is powerful and elegant, it is limited to simple-vs-simple testing problems. As soon as either hypothesis becomes composite, we must extend the theory.

5.3 Need for Generalized NP Lemma

In real statistical practice, hypotheses almost never specify a single parameter value. Instead, they usually specify **ranges** or **sets** of values. Examples:

$H0:\mu \leq \mu0$ vs $H1:\mu > \mu0$

$H0: p = p0$ vs $H1: p = p0$

$H0: \theta \geq 0$ vs $H1: \theta < 0$

In all these cases, the parameter under either H0 or H1 is not a single number. The NP Lemma cannot handle such situations, because:

1. Multiple Likelihood Ratios

For composite hypotheses, each value inside the parameter space gives a different likelihood ratio:

$$\Lambda_\theta(x) = \frac{f(x|\theta 0)}{f(x|\theta)}$$

which varies as θ.

2. Type I Error must be controlled for all values

To be a level-α test, the test must maintain the error bound **uniformly**:

$$\sup_{\theta \in \Theta 0} P\theta(\text{reject } H0) \leq \alpha$$

This is far stricter than the simple-case requirement.

3. Power must be maximized across a set

For composite alternatives, optimality requires:

$\sup_{\theta \in \Theta 1} P\theta(\text{reject } H0)$ to be as large as possible. A test that is powerful at one value of θ might not be powerful at another.

4. Optimal tests may not exist

Conflicting power requirements over ranges of parameters often make it impossible to find a single test that is best for all alternative values. This is why UMP tests rarely exist in two-sided or multi-parameter settings.

Conclusion

These difficulties lead naturally to the **Generalized NP Lemma**, which extends NP Lemma to composite hypotheses by using supremum and infimum power comparisons instead of pointwise comparisons.

5.4 STATEMENT AND INTERPRETATION OF GNP LEMMA (EXPANDED)

Let:

$\Theta 0$ = parameter space under H0

Θ1 = parameter space under H1

A test φ(x) is of level α if:

$$\sup_{\theta \in \Theta 0} E\theta[\varphi(X)] \leq \alpha$$

Generalized NP Lemma

A test φ∗is **Uniformly Most Powerful (UMP)** for testing H0 vs H1 if:

$$\inf_{\theta \in \Theta 1} E\theta[\varphi*(X)] \geq \inf_{\theta \in \Theta 1} E\theta[\varphi(X)]$$

for all level-α tests φ.

Interpretation

- The test must be valid (size ≤ α) **for every θ in Θ_0**.
- The power of φ∗ must dominate every other test **for every θ in Θ_1**.
- This ensures **uniform superiority**, not just pointwise superiority.

5.5 OPTIMALITY FOR COMPOSITE HYPOTHESES (EXPANDED)

To be optimal under composite hypotheses:

1. Size across Θ_0

$$\sup_{\theta \in \Theta 0} P\theta(\text{reject}) \leq \alpha$$

The worst-case Type I error determines validity.

2. Power Across Θ_1

$$\sup_{\theta \in \Theta 1} P\theta(\text{reject})$$

This ensures performance does not drop for any alternative.

3. Uniform Dominance

A test is optimal only if its power curve **never falls below** that of any competitor for any alternative parameter value. This strong requirement explains why UMP tests are rare unless additional structural properties (like MLR) exist.

5.6 APPLICATIONS AND EXAMPLES

Example 1: Normal Distribution (One-Sided Alternative)

Testing:

H0: $\mu \leq \mu 0$ vs H1: $\mu > \mu 0$..

The likelihood ratio is a monotone function of $\bar{x}$.
Thus, the UMP test is:

Reject H0 if $\bar{x} > c$.

Example 2: Binomial Parameter

Testing:

H0: $p \leq p0$ vs H1: $p > p0$

Reject H0 when X (number of successes) is large.

Example 3: Poisson Mean

Testing:

H0: $\lambda \leq \lambda 0$ vs H1: $\lambda > \lambda 0$

Reject for large values of X.

These examples demonstrate how the GNP Lemma guides the construction of UMP one-sided tests.

5.7 LIMITATIONS OF GNP LEMMA

Although the Generalized Neyman–Pearson Lemma provides a powerful theoretical framework for identifying optimal tests under composite hypotheses, its practical application is often limited by several important factors. These limitations help explain why UMP tests are relatively rare and why additional structural properties—such as monotone likelihood ratios—play a crucial role in modern hypothesis testing.

1. The GNP Lemma Does Not Directly Construct the Test

Unlike the classical NP Lemma, which gives a clear likelihood ratio rule for the rejection region, the GNP Lemma does **not specify an explicit decision rule**.
It tells us *what an optimal test must satisfy*, but it does not provide: a functional form of the test statistic, the exact threshold, or the structure of the rejection region.

As a result, the lemma is more of a **theoretical criterion** than a constructive tool. In many cases, solving the supremum/infimum inequalities required by the lemma is mathematically challenging.

2. UMP Tests rarely exist when testing two-sided alternatives

The GNP Lemma can guarantee existence of UMP tests only when the structure of the problem aligns perfectly with its conditions.

However, in a large number of important testing scenarios—particularly **two-sided tests** such as:

H0: $\theta = \theta 0$ vs H1: $\theta \neq \theta 0$, a UMP test **does not exist**.

This is because:

- Maximal power for detecting $\theta > \theta 0$ forces the rejection region toward one tail.
- Maximal power for detecting $\theta < \theta 0$ forces it toward the opposite tail.

Since no single rejection region can simultaneously dominate all others in both directions, uniform optimality fails.

Thus, the GNP Lemma reveals *why* UMP tests do not exist, but it cannot fix the problem on its own.

3. Supremum/Infimum over parameter spaces can be difficult to evaluate

The GNP Lemma requires evaluating the following two expressions:

The **supremum** of Type I error over $\Theta 0$

$$\sup_{\theta \in \Theta 0} E\theta[\varphi(X)]$$

The **infimum** of power over $\Theta 1$

$$\inf_{\theta \in \Theta 1} E\theta[\varphi(X)]$$

These quantities may be: computationally intensive, analytically intractable, or impossible to compute without numerical methods. In many real applications, these extremum values do not have neat closed-form solutions. This makes the use of GNP Lemma impractical unless the model has simple mathematical properties.

4. The Lemma needs additional structural properties to be useful

The GNP Lemma is most effective when used together with other structural properties such as:

(a) Monotone Likelihood Ratio (MLR)

MLR allows us to reduce the test to a rejection region based on an ordered statistic.
Without MLR, even if the GNP Lemma states that a UMP test exists, we may not know how to construct it.

(b) Exponential Family Structure

The one-parameter exponential family has:sufficient statistics, natural MLR properties. well-behaved likelihood ratios. These properties allow the GNP Lemma to provide meaningful

results. In general distributions lacking such structure, the lemma may not yield practical conclusions.

5. The GNP Lemma ensures optimality but not feasibility

A test that satisfies the GNP optimality criteria may: exist mathematically but not be easy to compute, depend on unknown parameters, require randomization to achieve exact level, or

be sensitive to model assumptions. Thus, while theoretically sound, GNP-based tests may not always be practical for real-world data analysis.

6. Does not address nuisance parameters

Many common problems in statistics involve nuisance parameters.
The GNP Lemma: does not provide guidance for eliminating or adjusting for nuisance parameters, cannot guarantee uniform optimality across multidimensional parameter spaces.

This significantly restricts its applicability to simple, well-structured models.

Summary of Limitations

- It **does not provide explicit test rules** like the classical NP Lemma.
- UMP tests derived using the GNP Lemma often **do not exist** for two-sided or multi-parameter alternatives.
- Requires solving **supremum/infimum optimizations**, which are mathematically demanding.
- The lemma is **useful only when combined** with additional conditions like MLR or exponential family properties.
- Does not naturally address **nuisance parameters** or **complex models**.
- Sometimes optimal tests exist mathematically but are **not practically implementable**.

5.8 CONCLUSION

The Generalized Neyman–Pearson Lemma broadens the scope of the classical NP Lemma by addressing the more realistic and commonly encountered case of **composite hypotheses**, where parameters vary over a range rather than being fixed at a single value. It establishes a rigorous theoretical foundation for identifying **Uniformly Most Powerful (UMP)** tests in situations involving **one-sided composite alternatives**, thereby extending optimality principles beyond simple-vs-simple comparisons.

However, the practical usefulness of the GNP Lemma is limited unless the underlying probability model possesses additional structural properties. In many settings, the lemma alone cannot guarantee the existence or explicit form of a UMP test. This is why further assumptions—such as those found in the **one-parameter exponential family** or in distributions exhibiting the **Monotone Likelihood Ratio (MLR)** property—become essential. These structural features help simplify the supremum/infimum comparisons required by the lemma and allow the construction of concrete, implementable tests.

In summary, while the GNP Lemma provides the conceptual framework for optimal testing under composite hypotheses, it typically requires the support of stronger model

characteristics—such as MLR or exponential family structure—to yield practical and fully specified UMP tests.

5.9 SELF-ASSESSMENT QUESTIONS

1. What is a composite hypothesis?
2. How does the classical NP Lemma differ from the GNP Lemma?
3. State the Generalized Neyman–Pearson Lemma in your own words.
4. What are the two main conditions an optimal test must satisfy under composite hypotheses?
5. Give an example of a UMP test derived using the GNP Lemma.
6. Why might UMP tests fail to exist in some composite testing scenarios?

5.10 SUGGESTED READINGS

1. Lehmann, E.L. & Romano, J. P. *(Testing Statistical Hypotheses)*
2. Casella, G. & Berger, R. L. *(Statistical Inference)*
3. Rao, C. R. *(Linear Statistical Inference and Its Applications)*
4. Mood, A. M., Graybill, F. A. & Boes, D. C. *(Introduction to the Theory of Statistics)*

Dr. M.Vijaya Lakshmi

LESSON -6

UMP TESTS FOR SIMPLE NULL

OBJECTIVES

- By the end of Lesson 6, students will be able to:
- Define Uniformly Most Powerful (UMP) tests.
- Use Neyman–Pearson (NP) and Generalized Neyman–Pearson (GNP) Lemmas to construct UMP tests for one-sided alternatives.
- Understand uniform power and why UMP tests are preferred when they exist.
- Derive UMP tests for simple null vs. one-sided alternatives using the likelihood ratio.
- Apply UMP principles to Normal, Binomial, and Poisson models.
- Recognize the conditions required for the existence of UMP tests, especially the role of monotone likelihood ratio (MLR).
- Identify scenarios where UMP tests fail to exist even for one-sided alternatives.

STRUCTURE

6.1 INTRODUCTION TO UMP TESTS

In hypothesis testing, we often compare a null hypothesis H0 with an alternative hypothesis H1. For a given significance level α (probability of Type I error), many different tests may satisfy the condition "size = α". However, not all of them perform equally well in detecting true departures from H0.

A **Uniformly Most Powerful (UMP) test** is the test that performs the *best* among all valid tests of size α.

What does "most powerful" mean?

The *power* of a test at a parameter value θ is:

$\beta(\theta)=P\theta(\text{Reject H0})$

Higher power means the test is better at detecting when the alternative hypothesis is true.

What does "uniformly" mean?

For a composite alternative (for example, H1:θ>θ0), there are many possible parameter values under H1. A test is **UMP** if it has the *largest* power **for every one** of those values of θ.

A test is UMP if β(θβ∗(θ)for all tests β∗, and for all θ∈H1..

So **uniformity** means:

No other test beats it anywhere in the entire alternative region.

Why NP Lemma is not enough

The **Neyman–Pearson Lemma** provides the most powerful test only when:

H0 is simple (only one value of θ), and H1 is simple (only one value of θ1).

But in many problems, the alternative has *many possible values*.
For example:

H0:μ = μ0, H1:μ > μ0

Here, μ can be 10.1, 10.5, 11, 15, etc. NP Lemma alone cannot tell us how to find a single test that is best for *all* these values. This is where UMP theory becomes important.

When UMP Tests Exist

UMP tests are not guaranteed to exist for all kinds of hypotheses.
They exist mainly when: The distributions belong to a **one-parameter exponential family**, and The family has the **Monotone Likelihood Ratio (MLR)** property.

Under these conditions, we can extend the NP result to composite alternatives (especially **one-sided** ones).

If a UMP test exists:

- It is always the best choice because no other test can outperform it.
- It simplifies testing: the decision rule is clear and optimal.
- It avoids searching through many possible tests.

However, UMP tests do **not** generally exist for: Two-sided alternatives, Multi-parameter models, Models without MLR. These limitations explain why this topic is important and why statisticians sometimes use UMPU or Likelihood Ratio Tests instead.

Example

Consider testing: H0: p = 0.3, H1: p > 0.3

using a Binomial model.

Because the binomial family has MLR in the number of successes X: The test that rejects H0 for **large values of** X turns out to be UMP for all $p > 0.3$. This means: No other size-α test gives higher power for any $p>0.3$. The test is best everywhere in the one-sided alternative.

6.2 ONE-SIDED ALTERNATIVES: MOTIVATION & SETUP — EXPANDED EXPLANATION

In many real-world situations, we are interested in detecting a change in **one particular direction**.
For example:

- Has the average weight of a product **increased** after a new machine was installed?
- Has the defect probability of a process **exceeded** the acceptable limit?
- Has the rate of accidents **increased** after a policy change?
- Is a new fertilizer causing **higher** crop yield?

In these cases, we don't care if the parameter is lower—we are specifically interested in whether it has become **larger**. This leads to **one-sided hypothesis testing**.

Typical One-Sided Hypothesis Setup
H0: $\theta = \theta 0$ vs. H1: $\theta > \theta 0$or H0: $\theta = \theta 0$ vs. H1: $\theta < \theta 0$

Here the alternative hypothesis includes **many** possible values of the parameter.

Example:
If H1: $\mu > 50$ can be 51, 55, 70, or any value greater than 50. This is a **composite alternative**.

Why One-Sided Tests Are Easier for UMP

For one-sided alternatives, the parameter moves **in only one direction**, and this often results in: A **clear ordering** in the likelihood ratio. A natural test statistic (like sample mean, number of successes). A rejection rule of the form:

$$T(X) > c \text{ or } T(X) < c$$

If the distribution has the **Monotone Likelihood Ratio (MLR)** property, then the model behaves in a predictable way: higher values of the statistic correspond to higher parameter values.
This allows us to use a *single threshold* to reject H0.

This forms the mathematical basis for the existence of **UMP tests** in one-sided cases.

Practical Interpretation

A one-sided UMP test tells us:

> "Whenever the parameter increases beyond the null value, this test is the most sensitive and effective detector—better than all other size-α tests."

This uniform superiority is what makes UMP tests so valuable in one-sided problems.

Simple Real-Life Example

Quality Control Problem:
A machine is supposed to produce bolts with average length μ=10\mu = 10μ=10 cm.

If the company wants to check **only whether the machine is producing longer bolts**, the test is:

H0: μ = 10 Vs H1:μ > 10

A sample mean greater than expected would lead to rejecting H0.
Since the normal distribution (with known variance) has MLR in the sample mean, the
Numerical Example

A sample of 25 items has an average weight of 102 grams.
The historical mean is 100 grams, and the standard deviation is known to be 8 grams.

To test:

H0:μ = 100 vs H1: μ > 100

Compute:

$$Z = \frac{(102-1008)}{\frac{8}{\sqrt{25}}25} = \frac{21}{6} = 1.25$$

If $z_{0.05} = 1.645$, then 1.25 < 1.645, so **do not reject** H0.

This test is the UMP test for detecting an increase in the mean.

6.3 UMP TESTS FOR SIMPLE NULL:

In this section, we explore **how UMP tests arise from the Neyman–Pearson framework** when the null hypothesis is simple and the alternative is one-sided. The starting point is the **Neyman–Pearson Lemma (NP Lemma)**, which states: For testing a simple null against a simple alternative, the most powerful test of size α is based on the likelihood ratio. Although NP Lemma only deals with **simple vs simple**, it lays the foundation for constructing UMP tests when the alternative is **one-sided and composite**.

1. NP Lemma for Simple Hypotheses

Consider:

H0: θ = θ0, H1: θ = θ1 > θ0

Let the likelihoods be:

L0(x) = f(x;θ0)

$L1(x) = f(x;\theta 1)$

The NP Lemma says:

Reject H0 if $\frac{L1(x)}{L0(x)} > k$

for some constant k chosen to ensure size α.

This is the **most powerful** test for distinguishing θ0 from a single alternative value θ1.

2. Moving to a One-Sided Composite Alternative

Now consider:

H0: $\theta = \theta 0$, H1 : $\theta > \theta 0$

Here, the alternative includes infinitely many values: θ1, θ2, θ3,… where each $\theta i > \theta 0$. A UMP test must be **most powerful simultaneously** against *all* these values.

This is only possible if the likelihood ratio:

$$\frac{f(x;\theta)}{f(x;\theta 0)}$$

changes in a **consistent direction** as θ increases.

This property is known as the **Monotone Likelihood Ratio (MLR)**.

3. The Role of Monotone Likelihood Ratio (MLR)

A family of distributions has MLR in a statistic T(X) if:

$\frac{f(x;\theta 1)}{f(x;\theta 0)}$ increases as T(X) increases whenever $\theta 1 > \theta 0$..

This means larger values of T(X) provide stronger evidence against H0.

Examples of statistics with MLR:

- Sample mean in Normal distribution
- Number of successes in Binomial distribution
- Count in Poisson distribution

When MLR holds, the NP test for each $\theta 1 > \theta 0$ has **the same rejection region shape**:

$T(X) > cT(X) > cT(X) > c$

so a **single** test works best for *all* one-sided alternatives.

This test becomes the **UMP test**.

4. UMP Test Form for Simple Null vs One-Sided Alternatives

If the distribution has MLR in T(X), the UMP test for:

H0: $\theta = \theta 0$ vs H1: $\theta > \theta 0$

is:

Reject H0 if $T(X) > c\alpha$

where $c\alpha$ is chosen so that:

$P\theta 0(T(X) > c\alpha) = \alpha$.

This test is:

Simple to compute

Based on an intuitively increasing statistic

Uniformly most powerful for all $\theta > \theta 0$

5. Example: Binomial Model

Let $X \sim Bin(n,p)$.

Test:

H0: $p = 0.3$, H1: $p > 0.3$

The likelihood ratio:

$\frac{f(x;p1)}{f(x;0.3)}$ is increasing in x.Thus, MLR holds, and the UMP test becomes:

Reject H0 if $X \geq c$ for a suitable cutoff c.

This test is UMP because higher numbers of successes are more consistent with higher p.

6. Numerical Example (Simple)

Suppose $X \sim Poisson(\lambda)$.

Test:

H0: $\lambda = 3$ vs H1: $\lambda > 3$

Since Poisson distribution has MLR in X, the UMP test is:

Reject H0 if $X \geq c$

For $\alpha=0.05$:

Find ccc such that:

$P(X \geq c|\lambda=3) = 0.05$

Using values:

$P(X \geq 7) = 0.033$

$P(X \geq 6) = 0.084$

So c = 7.

Test rule:

Reject H0 **if number of events ≥ 7.** This is the UMP test.

6.4 LIKELIHOOD RATIO ORDERING AND REJECTION REGION

To construct a UMP test for a one-sided alternative, the crucial idea is to understand **how the likelihood ratio behaves as the data changes**. If the likelihood ratio increases (or decreases) in a predictable way with a statistic T(X), then we can form a simple and optimal rejection region. This behaviour is captured by the **Monotone Likelihood Ratio (MLR)** property.

1. Likelihood Ratio for a Simple Null vs. Composite One-Sided Alternative

Suppose we test:

H0: $\theta = \theta 0$ vs H1: $\theta > \theta 0$.

For any $\theta 1 > \theta 0$, consider the likelihood ratio:

$$\Lambda(x) = \frac{f(x; \theta 1)}{f(x; \theta 0)}$$

If the family has MLR in some statistic T(X), then:

- When the parameter increases,
- The likelihood ratio also tends to increase whenever T(X) increases.

This gives a **natural order** for deciding when data provides stronger evidence against H0.

2. The MLR Property and Ordering of Evidence

A family $\{f(x;\theta)\}$ has MLR in T(X) if:

$\frac{f(x;\theta 1)}{f(x;\theta 0)}$ is increasing in T(X)whenever θ1 > θ0.

Interpretation:

- Higher values of T(X) make the alternative more likely relative to the null.
- Lower values of T(X) make the null more likely.

So *evidence* accumulates along a single direction.

This allows us to use a **single cut-off** in T(X) to define a test.

3. Form of the UMP Rejection Region

If MLR holds, the UMP test for H1:θ > θ0 always has the form:

Reject H0 if T(X) > cα

where the constant c is chosen such that:

Pθ0(T(X) > cα) = α.

No other test of size α **can have higher power for all** θ>θ0**.**

This simple threshold rule is a hallmark of UMP tests.

4. Why This Rejection Region is Optimal

Because of the MLR property:

For any fixed alternative value θ1, the likelihood ratio is highest when T(X) is large.

The Neyman–Pearson test for H1:θ=θ1H_1:\theta=\theta_1H1:θ=θ1 is:

- Reject H0 if T(X) is large.

Since the same direction works for **every** θ1 > θ0, a single test works against all one-sided alternatives.

Thus, the rejection region is both:

- **Simple**
- **Uniformly Most Powerful**

5. Illustrative Example (Binomial)

Let X ∼Bin(n,p).

Test:

H0: p = p0 vs. H1: p > p0.

The likelihood ratio:

$\Lambda(x) = \left(\frac{p1}{p0}\right)^{x} \left(\frac{1-p1}{1-p0}\right)^{n-x}$

is **increasing in** x because the exponent of $\frac{p1}{p0}$ is x and p1 > p0.

Thus, the UMP test is:

Reject H0 if X ≥ c.

6. Illustrative Example (Normal Mean)

Suppose X1,X2,…,Xn are i.i.d. $N(\mu,\sigma^2)$

Test:

H0: μ = μ0 vs H1: μ > μ0.

The likelihood ratio is monotone in the sample mean $\bar{x}$.

Thus, the UMP test is:

Reject H0 if $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} -> Z_{\alpha}$. Reject H0

Again, we get a threshold-based test.

Summary

When the model has **MLR**, the likelihood ratio increases or decreases in a predictable way with a statistic T(X). This leads to a **single ordered rejection region** of the form T(X)>c. This region works simultaneously for all one-sided alternatives, giving a **UMP test**. Many common distributions (Normal, Binomial, Poisson) have MLR structure.

6.5 EXAMPLES: NORMAL, BINOMIAL, POISSON

To understand how UMP tests work in practice, it is helpful to see how they arise in familiar statistical models.The three most common one-parameter families—**Normal**, **Binomial**, and **Poisson**—all possess the **Monotone Likelihood Ratio (MLR)** property. As a result, UMP tests for one-sided alternatives exist in each case. Below, we examine each model separately and derive the UMP test.

A. UMP Test in the Normal Distribution (Variance Known)

Consider independent observations

$X1, X2, \ldots, Xn \sim iidN(\mu, \sigma^2)$

where the variance σ^2 is known.

We want to test:

H0: $\mu = \mu 0$ vs H1: $\mu > \mu 0$..

1. Likelihood Ratio

The likelihood ratio for two means $\mu 1 > \mu 0$ is **increasing in the sample mean** $\bar{x}$. Thus, the Normal family has MLR in $\bar{x}$.

2. UMP Test Form

Reject H0 if $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow Z_\alpha$ {Reject }

This is the classical **one-sample Z-test** for a one-sided alternative.

3. Simple Numerical Example

A machine produces items with known standard deviation $\sigma = 4$. Sample size n=25, sample mean $\bar{x} = 52$.

Test:

H0:$\mu = 50$ vs H1: $\mu > 50$

Compute:

$$Z = \frac{(52-50)}{\frac{4}{\sqrt{25}}} = \frac{2}{0.8} = 2.5$$

Critical value: $z_{0.05} = 1.645$.

Because 2.5 >1.645, we **reject** H0. This is the **UMP test** for this setup.

B. UMP Test in the Binomial Distribution

Let:

$X \sim Bin(n,p)$

Test:

H0:p = p0 vs H1: p > p0..

1. Likelihood Ratio Behavior

The likelihood ratio

$$\frac{f(x;p1)}{f(x;p0)} = \left(\frac{p1}{p0}\right)^{x} \left(\frac{1-p1}{1-p0}\right)^{n-x}$$ is **increasing in** x when p1 > p0 .

Thus, larger values of X support the alternative more strongly.

2. UMP Test Form
Reject H0 if $X \geq c$

where c is chosen so that:

$P(X \geq c|p = p0) = \alpha$.

This is a **right-tail rejection region**.

3. Simple Numerical Example

Suppose a production process has defect probability p0 = 0.10. A sample of n = 20 items shows 5 defectives.

Test:

H0: p = 0.10 vs H1: p > 0.10

Compute:

p-value = $P(X \geq 5|p = 0.10)$

Using binomial table:

$P(X \geq 5) \approx 0.044$.

Since 0.044 < 0.05, we **reject** H0.

This test is the **UMP test** for detecting an increase in the defect rate.

C. UMP Test in the Poisson Distribution

Let:

$X \sim Poisson(\lambda)$

Test:

H0: $\lambda = \lambda 0$ vs H1: $\lambda > \lambda 0$.

1. Likelihood Ratio

The likelihood ratio for Poisson is:

$$\frac{f(x;;\lambda 1)}{f(x;\lambda 0)} = \left(\frac{\lambda 1}{\lambda 0}\right)^{x} (e)^{-(\lambda 1-\lambda 0)}$$

Because the expression depends on x through a term of the form $(a)^{x}$, it is **increasing in** x when $\lambda 1 > \lambda 0$.

Thus, the family has MLR in X.

2. UMP Test Form
Reject H0 if $X \geq c$. The cutoff c is selected so that: $P(X \geq c|\lambda = \lambda 0) = \alpha$.

3. Simple Numerical Example

A call center expects $\lambda 0 = 5$ calls per hour. After an advertising campaign, they observe X=9 calls in an hour.

Test:

H0: $\lambda = 5$ vs H1: $\lambda > 5$.

Find:

p-value = $P(X \geq 9|\lambda = 5)$.

Using Poisson table:

$P(X \geq 9) \approx 0.048$.

Since 0.048<0.05, we **reject** H0.

This is the **UMP test** for an increased call rate.

Summary

Normal, **Binomial**, and **Poisson** distributions all have **MLR**, so UMP tests exist for one-sided alternatives.

The UMP rejection rule is always a **right-tail test** of the form:

$T(X) > c$

Thresholds depend on the distribution and significance level. These tests are *provably* optimal for detecting parameter increases.

6.6 POWER FUNCTION AND UNIFORMITY

The **power function** tells us how effective a test is at detecting departures from the null hypothesis.
Understanding this concept is essential because the definition of a **Uniformly Most Powerful**

(UMP) test is based entirely on how the power function behaves across different parameter values.

1. What is the Power Function?

For a test of H0 against H1, the **power function** is:

$\beta(\theta) = P\theta(\text{Reject H0})$

This represents:

The probability that the test correctly rejects the null hypothesis

When the true parameter value is θ

At the null value:
$\beta(\theta 0) = \alpha$ because the test is constructed to have size α.

Inside the alternative:

We want $\beta(\theta)$to be as **large as possible** for every $\theta \in H1$.

2. How Power Behaves in One-Sided Tests

In one-sided tests (e.g., $H1: \theta > \theta 0$): As θ increases above $\theta 0$, The probability of rejecting H0 also increases. Thus, a "good" test will have a power function that rises quickly.

For example, in a test of $H0: \mu = 10$ Vs $H1: \mu > 10$

$\beta(10) = \alpha$, $\beta(12) > \beta(11) > \beta(10)$

This means the test gets stronger as the true parameter moves further into the alternative.

3. Uniformity in UMP Tests

A test is **Uniformly Most Powerful** if:

$\beta(\theta) \geq \beta*(\theta)$for every $\theta \in H1$,

for *every* other test $\beta*$of the same size.

Uniformity means:

- The test is best **not only at one value**,
- But at **all** values in the alternative hypothesis.

No competing test beats it at any point. Most tests may perform well for some parameter values but poorly for others. A UMP test performs **better or equal everywhere**.

This makes UMP tests the "gold standard" whenever they exist.

4. Power Function Example: Normal Mean Test

Consider:

$X1,\ldots,Xn \sim N(\mu,\sigma^2)$

Testing:

$H0:\mu = \mu0$ vs $H1:\mu > \mu0$

The UMP test is:

Reject H0 if $Z = Z = \frac{\bar{x}-\mu0}{\frac{\sigma}{\sqrt{n}}} \rightarrow Z_{\alpha}$.

Power Function Derivation

Under true mean μ:

$Z \sim N(\frac{\mu-\mu0}{\frac{\sigma}{\sqrt{n}}}, 1)$

Power is:

$\beta(\mu)=P(Z > z\alpha|\mu) = 1-\Phi(z\alpha - \frac{\mu-\mu0}{\frac{\sigma}{\sqrt{n}}})$,

where Φ is the standard normal CDF.

Behavior:

- When $\mu = \mu0$: $\beta(\mu) = \alpha$
- As μ increases, $\beta(\mu)$ increases
- As $\mu\rightarrow\infty$, $\beta(\mu)\rightarrow1$

This rising curve confirms high sensitivity for detecting increases in μ.

5. Power Function Example: Binomial Test

Suppose:

$X\sim Bin(20,p)$

Test:

$H0:p = 0.3$ vs $H1:p > 0.3$

UMP test: Reject H0 if $X \geq 9$.

Compute power at different p:

p	Power $\beta(p)=P(X \geq 9)$
0.30	0.05 (size)
0.40	0.21
0.50	0.53
0.60	0.80

This monotonic increase shows uniform improvement as p increases.

6. Why Power & Uniformity Matter

Understanding power is essential because:

- **A test with low power is practically useless** even if it has correct size.
- UMP tests guarantee the **best performance** for detecting changes in the direction of interest.
- Uniformity ensures no other test outperforms it **anywhere** in the alternative region.

Thus, UMP tests, when available, are the optimal choice for one-sided hypothesis testing.

6.7 NONEXISTENCE CASES FOR ONE-SIDED TESTS

Although UMP tests exist for many common one-parameter families (Normal, Binomial, Poisson), there are important situations where **UMP tests do not exist**, even for one-sided alternatives.
Understanding *why* they fail is essential because it shows the limitations of the Neyman–Pearson approach and motivates more general testing methods (like UMPU and LRTs).

1. When UMP Tests Fail to Exist

UMP tests may fail to exist when:

a) The model does not have the MLR property

UMP tests rely heavily on the fact that the likelihood ratio can be ordered using a simple statistic T(X). If the distribution **does not exhibit MLR**, then: The likelihood ratio may increase for some values of X and decrease for others. This means we cannot find a **single direction** in which evidence for H1 grows, so no single test works best for all $\theta > \theta 0$.

b) The distribution depends on more than one parameter

In multi-parameter settings (e.g., Normal with unknown mean and unknown variance):

$X1,\ldots,Xn \sim N(\mu,\sigma^2)$

Test:

$H0:\mu = \mu0$ vs $H1: \mu > \mu0$

But since σ^2 is unknown:

- The likelihood depends on both μ and σ^2 .
- We cannot order the likelihood ratio cleanly
- Thus, UMP test **does not exist**

This is why we use the **t-test**, which is optimal in a different sense (UMPU under symmetry), not UMP.

c) The test statistic has a complicated or non-monotonic likelihood ratio

Sometimes the LR behaves irregularly; This prevents identifying a single cutoff region. Such cases occur in distributions like: Gamma with unknown shape, Non-standard mixture models, Non-exponential family distributions, Without MLR, UMP tests generally do not exist.

2. Example where ump does not exist (normal with unknown variance)

Consider:

$H0:\mu = \mu0$ vs $H1: \mu > \mu0$

but σ^2 is unknown.

The likelihood ratio involves both μ and σ^2. The test that is most powerful for one value of σ^2 might not be most powerful for another.

Conclusion:
There is *no single test* that dominates all others for all σ^2.

Thus, **no UMP test exists**. Instead, the classical **t-test** is used, which is UMP **unbiased**, not UMP overall.

3. Example where ump does not exist (gamma distribution)

Suppose:

$X \sim Gamma(\alpha,\beta)$

If we test the rate or shape parameter and the other is unknown, the likelihood ratio cannot be arranged in a monotone way. Thus, **UMP tests fail**, and instead likelihood ratio tests or Rao score tests are used.

4. Key points

A UMP test for one-sided alternatives exists only when: The distribution belongs to a **one-parameter exponential family**, and There is a statistic T(X) with **monotone likelihood**

ratio (MLR). If **either** condition fails: Likelihood ratios cannot be consistently ordered. Multiple competing tests will dominate different parts of the alternative. Therefore, **no UMP test is possible.** This explains why UMP tests are powerful but rare, and why we rely on other methods in many practical situations.

6.8 CONCLUSION

In this lesson, we developed a complete understanding of **Uniformly Most Powerful (UMP) tests**, their construction, and their limitations. The idea of a UMP test is central in hypothesis testing because it represents the **best possible test** for detecting a change in a parameter in **one specific direction**. A UMP test is one that has the **highest power** among all tests of a given size αfor **every** parameter value in the alternative hypothesis. This is a very strong requirement. For a test to be UMP, it must: Control Type I error at level α, and Perform **better than every other test** across the entire alternative region. This is why UMP tests are rare—they demand consistent superiority everywhere.

We found that UMP tests **do** exist in some very important situations:

a) One-parameter exponential families

Examples include:

- Normal distribution with known variance
- Binomial distribution
- Poisson distribution

These families have simple forms and are mathematically well-behaved.

b) When there is a Monotone Likelihood Ratio (MLR)

If the likelihood ratio increases or decreases in a predictable manner with a statistic T(X), then a simple cutoff rule:

Reject H0 if T(X) > c, becomes optimal for all values of the alternative.

This is what allows the existence of UMP tests for **one-sided alternatives**.

3. When UMP Tests Do ***Not*** Exist

Despite their advantages, UMP tests do **not** exist in many realistic situations:

- Two-sided alternatives
- Multiple unknown parameters (e.g., unknown variance)
- Distributions without MLR
- Complicated or irregular likelihood behavior

In such cases, no single test can dominate all others uniformly. This naturally leads to alternative test concepts, such as:

- **UMPU** (Uniformly Most Powerful Unbiased) tests

- **Likelihood Ratio Tests (LRT)**
- **Wald and Score tests**

These methods fill the gap when UMP tests are not available.

4. Importantance UMP Tests

Even though UMP tests are limited, they are important because:

They give **clear, simple rejection rules** (usually based on one statistic).

They provide **optimal performance** for detecting directional changes.

They serve as the theoretical foundation for many standard tests.

For example:

- The one-sided **Z-test** for a mean
- The **Binomial right-tail test** for proportions
- The **Poisson count test**

are all UMP tests.

This makes them extremely useful in practical applications across engineering, agriculture, medicine, manufacturing, and economics.

Summary:

- UMP tests are powerful, elegant solutions when the statistical model is simple, one-parameter, and well-structured.
- Their existence relies heavily on the MLR property.
- When these conditions hold, UMP tests give us the *best possible decision rule* for one-sided hypotheses.
- When they don't hold, we must move to more general testing ideas like UMPU or LRTs.

Thus, UMP tests are both a **cornerstone** and a **limit case** of classical hypothesis testing theory.

6.9 SELF-ASSESSMENT QUESTIONS

1. Define a Uniformly Most Powerful (UMP) test. How does the Neyman–Pearson Lemma relate to UMP tests?
2. What is the Monotone Likelihood Ratio (MLR) property?
3. Explain why UMP tests typically exist only for one-sided alternatives.
4. Why do UMP tests fail when the variance in a Normal distribution is unknown?
5. How is the rejection region of a UMP test typically structured?
6. A sample of size 25 is taken from a Normal distribution with known variance $\sigma = 6$. You want to test $H0:\mu = 50$ vs $H1:\mu > 50$. The sample mean is 53. At $\alpha=0.05$, determine whether to reject H0. *(Hint: compute a Z-value)*

7. A Binomial random variable X∼Bin(20,p)X \sim \text{Bin}(20, p)X∼Bin(20,p) is used to test: H0: p = 0.3 vs H1: p > 0.3. Suppose you observe X = 8. Compute the p-value.
8. A Poisson random variable X has mean λ. Test H0:λ = 4 vs H1:λ > 4. If X = 9, estimate the p-value.

6.10 SUGGESTED READINGS

· **Testing Statistical Hypotheses** — E. L. Lehmann & J. P. Romano

· **Statistical Inference** — George Casella & Roger L. Berger

· **Introduction to the Theory of Statistics** — Mood, Graybill & Boes

· **Mathematical Statistics with Applications** — Wackerly, Mendenhall & Scheaffer

· **Theory of Point Estimation & Testing** — Hogg, McKean & Craig

Dr. M.Vijaya Lakshmi

LESSON -7

TWO-SIDED ALTERNATIVES & UMP LIMITS

OBJECTIVES

By the end of Lesson 7, students will be able to:

- Explain UMP tests in the one-parameter exponential family.
- Describe the Monotone Likelihood Ratio (MLR) property and its role in constructing UMP tests.
- Apply the Karlin–Rubin theorem to derive UMP tests for one-sided hypotheses.
- Extend UMP results from exponential family models to general MLR distributions.
- Understand why UMP tests do not exist for simple null vs. two-sided alternatives.
- Distinguish between UMP, UMPU, and LRT when UMP does not exist.
- Analyze examples illustrating existence and nonexistence of UMP tests.

STRUCTURE

7.1 INTRODUCTION OF EXPONENTIAL FAMILY AND MLR

Many widely used probability models—such as the **Normal distribution with known variance**, **Binomial**, **Poisson**, and **Exponential** distributions—can be written in a special mathematical form called the **one-parameter exponential family**. This form is important because it automatically gives us: a **natural sufficient statistic** T(X), and a structured way to study how evidence changes as the data changes.

1. Why Exponential Families Matter for UMP Tests

A distribution belongs to the one-parameter exponential family if it can be written as:

$f(x;\theta)=h(x)\exp[\eta(\theta)T(x)-A(\theta)]$.

Here:

- T(X) is a statistic that captures all the information about the parameter θ;
- η(θ)\ is the *natural parameter*;
- A(θ) ensures the density integrates (or sums) to 1.

This structure makes the behavior of the likelihood ratio much easier to analyze.

2. Understanding the Monotone Likelihood Ratio (MLR)

The **Monotone Likelihood Ratio (MLR)** property means:

$\frac{f(x;\theta 1)}{f(x;\theta 0)}$ is increasing in T(X)whenever $\theta 1>\theta 0$.

Interpretation:
As T(X) increases, the data becomes **more supportive of a larger parameter value**. This "one-directional" change is crucial because it lets us order data from **least to most favorable** for the alternative.

3. Why MLR Leads to UMP Tests

If the likelihood ratio always rises as T(X) rises, then rejecting H0 for **large** values of T(X)is a uniformly optimal strategy. Thus, when MLR holds: Reject H0 if T(X) > c.This gives the **UMP test** for one-sided alternatives.

4. Examples: How MLR Looks in Common Distributions

(a) Normal Distribution (Known Variance)

For $X1,\ldots,Xn\sim N(\mu,\sigma^2)$: Sufficient statistic: $T(X) = \bar{X}$ is MLR in $\bar{X}$. If the true mean increases above μ0, the likelihood ratio increases continuously with $\bar{X}$. So large sample means give stronger evidence against H0. This yields the UMP test:

Reject H0 if $\bar{X} > \mu_0+z_\alpha\ \sigma_n$..

(b) Binomial Distribution

Let X∼Bin(n,p).

Likelihood ratio for p1 > p0:

$$\frac{f(x;p1)}{f(x;p0)} = \left(\frac{p1}{p0}\right)^{x} \left(\frac{1-p1}{1-p0}\right)^{n-x}$$

This clearly **increases with** x. Thus, the number of successes T(X)=X is the ordering statistic.

UMP test:

Reject H0 if $X \geq c$.

(c) Poisson Distribution

If $X \sim Poisson(\lambda)$:

$$\frac{f(x;;\lambda 1)}{f(x;\lambda 0)} = \left(\frac{\lambda 1}{\lambda 0}\right)^{x} (e)^{-(\lambda 1-\lambda 0)}$$

For $\lambda 1>\lambda 0$, $\Lambda(x)$ increases in x.
So large counts provide stronger evidence for higher λ.

UMP test:

Reject H0 if $X \geq c$.

5. Why This Matters for Testing

Because exponential families have MLR, they are the **main class of models** where we can systematically construct UMP tests.

Whenever MLR holds:

- Evidence increases consistently with T(X).
- No data "contradicts itself" in different regions.
- The rejection region is always a **simple threshold**.
- The UMP test exists and is easy to compute.

This is why many classical tests—Z-test, t-test when variance is known, Binomial test, Poisson test—are UMP for one-sided alternatives.

7.2 UMP TESTS FOR ONE-SIDED NULL VS ONE-SIDED ALTERNATIVES

One of the main strengths of exponential families with MLR is that they allow us to build **Uniformly Most Powerful (UMP)** tests for **one-sided alternatives**. These are among the few situations in statistics where a test can be shown to be *optimal for all parameter values* in the alternative region.

1. Structure of One-Sided Hypotheses

A one-sided test focuses on detecting a **change in a specific direction**:

$H0: \theta \leq \theta 0$ vs. $H1: \theta > \theta 0$.

Here:

- H0 allows values up to $\theta 0$.
- H1includes **all values greater than** $\theta 0$.
- The direction of the test is clearly defined ("greater than").

This type of hypothesis is common in real applications, e.g.: Checking if machine output has **increased**. Testing if a treatment **improves** recovery time. Determining if a defect rate has **exceeded** a quality threshold.

2. Extending NP Lemma to one-sided composite alternatives

The Neyman–Pearson Lemma (NP Lemma) gives the **most powerful test** for:

$H0: \theta = \theta 0$ vs. $H1: \theta = \theta 1$, a **simple vs simple** comparison.

However, in one-sided problems, the alternative is **composite** (many possible θ).
To find a test that is most powerful for *all* $\theta > \theta 0$, we need:

A statistic T(X) that orders the data, and

Likelihood ratios that move consistently with T(X).

This is exactly what the **MLR property** provides.

3. The Key Role of MLR in UMP Tests

When the family has MLR in T(X), then for any $\theta 1 > \theta 0$:

$\frac{f(x;\theta 1)}{f(x;\theta 0)}$ increases as T(X) increases.

This means:

- Larger values of T(X) give stronger evidence for $\theta > \theta 0$.
- The NP rejection region for **any** $\theta 1 > \theta 0$ has the same form.
- A threshold rule $T(X) > c$ works for all alternatives.

Thus, the UMP test is guaranteed to exist.

4. Form of the UMP Test

For the one-sided alternative $H1: \theta > \theta 0$, the UMP test always takes the form:

Reject H0 if $T(X) > c\alpha$, Reject H0 if $T(X) > c\alpha$,

where the constant $c\alpha$ is chosen so that:

$P\theta 0(T(X) > c\alpha) = \alpha$.

This gives a **size-α** test with **uniformly maximum power**.

5. Real Examples of UMP One-Sided Tests

Example 1: Normal Distribution (Mean Test)
$X1,\ldots,Xn \sim N(\mu,\sigma^2)$,$\sigma^2$ known.,

Test: $H0:\mu \leq \mu 0$ Vs $H1:\mu>\mu 0$.

MLR in $T(X) = \bar{X}$ ensures a UMP test:

$$Z = \frac{\bar{X}-\mu 0}{\frac{\sigma}{\sqrt{n}}}.$$

Reject H0 if $Z> z\alpha$..

Interpretation:
Large sample means signal a larger μ.

Example 2: Binomial Distribution

Suppose $X\sim Bin(n,p)$.

Test:

$H0:p \leq p0$ Vs $H1:p > p0$..

Since MLR holds in $T(X)=X$:

UMP test:

Reject H0 if $X \geq c$.

Interpretation:
More successes indicate larger p.

Example 3: Poisson Distribution

If $X\sim Poisson(\lambda)$:

Test:

$H0:\lambda \leq \lambda 0$ Vs $H1:\lambda > \lambda 0$..

UMP test:

Reject H0 if $X \geq c$.

Interpretation:
Larger counts suggest a larger rate λ.

6. Why UMP Tests are so useful here

- For one-sided hypotheses in exponential families:
- The test is simple and easy to implement.
- The decision rule follows clear logic: **large** T(X) **means evidence for** H1
- The test is *provably the best* for detecting increases in the parameter.
- No other test, regardless of form, has better power for every θ>θ0.
- This makes UMP tests extremely valuable in practice.

7.3 CONSTRUCTION OF UMP TESTS UNDER MLR PROPERTY

Once we know that a distribution has the **Monotone Likelihood Ratio (MLR)** property in a statistic T(X), constructing a UMP test for a one-sided hypothesis becomes systematic and straightforward. The MLR ensures that **larger values of** T(X) **always support larger values of the parameter** θ. This provides a natural ordering of evidence and guarantees the existence of a uniformly most powerful test.

1. Using the Likelihood Ratio to Form the Test

For testing:

H0:θ ≤ θ0 vs H1:θ > θ0, consider any point θ1>θ0.
The **likelihood ratio** is:

$$\Lambda(x) = \frac{f(x;\theta 1)}{f(x;\theta 0)}.$$

If the family has **MLR in** T(X), then:

- Λ(x) increases whenever T(X) increases;
- All NP most powerful tests for different θ1's reject for **large values of** T(X);
- The rejection regions have the same shape.

Thus, one single test works for all alternatives θ>θ0\theta>\theta_0θ>θ0.

2. The General Form of the UMP Test

Because of MLR, the UMP test is always of the form:

Reject H0, if T(X) > cα

The cutoff cα:

Pθ0(T(X) > cα)= α..

- This construction ensures:
- Size α (correct Type I error),

- Maximum power for all $\theta>\theta 0$,
- A simple, threshold-based rule.

3. Why this Construction Works

MLR means:

- As T(X)gets larger, the distribution under $\theta 1 > \theta 0$ becomes more likely relative to $\theta 0$.
- Therefore, large values of T(X) correspond to stronger evidence against H0.

This justifies:

- Rejecting for large values of T(X),
- Using a single cutoff for the entire one-sided alternative.
- This makes UMP tests very easy to construct in exponential families.

4. Step-by-Step Construction Procedure

Identify the statistic T(X) in which the family has MLR.

Find the distribution of T(X) under $H0:\theta = \theta 0$.

Determine the cutoff $c\alpha$ such that

1. $P\theta 0(T(X) > c\alpha) = \alpha$.

Define the test:

2. Reject H0 if $T(X) > c\alpha$.\

State that this test is UMP against $H1:\theta>\theta 0$ due to the MLR property.

5. Examples

(a) Normal Distribution (Known Variance)
$X1,\ldots,Xn \sim N(\mu,\sigma^2)$. MLR in $T(X) = \bar{X}$.

Test:

$H0:\mu \leq \mu 0$ Vs $H1:\mu > \mu 0$

Construct UMP test:

Statistic: $T(X) = \bar{X}$

Under H0:

1. $\bar{X} \sim N(\mu 0,\sigma^2/n)$

Cutoff:

2. $C\alpha = \mu 0 + Z\alpha\sigma_n$

Rejection rule:

3. Reject H0 if $\bar{X} > \mu 0 + Z\alpha\ \sigma_n$

This is the classical **Z-test**, and it is UMP for the one-sided alternative.

(b) Binomial Distribution

Let X~Bin(n,p).

MLR in the statistic T(X)=X.

Test:

H0:p ≤ p0 Vs H1:p > p0

Construction:

Statistic: T(X)=X

Under H0:

1. X ~Bin(n,p0)

Find cα such that:

2. Pp0(X ≥ cα) = α

Rejection rule:

3. Reject H0 if X ≥ cα..

This is the UMP test for increased proportion.

(c) Poisson Distribution

Let X~Poisson(λ).

MLR in T(X)=X.

Test:

H0:λ ≤ λ0 Vs H1:λ>λ0

Construction:

Under H0:

1. $X \sim \text{Poisson}(\lambda 0)$

Cutoff $c\alpha$ solves

2. $P\lambda 0(X \geq c\alpha) = \alpha$

Rejection region:

3. $X \geq c\alpha$

Again, a simple and optimal test.

7.4 EXTENSION TO GENERAL DISTRIBUTIONS WITH MLR

Although exponential families provide the most natural setting for UMP tests, the existence of a **Uniformly Most Powerful (UMP)** test does **not** depend on whether a distribution belongs to the exponential family. What actually matters is whether the distribution has the **Monotone Likelihood Ratio (MLR)** property in some statistic T(X). If MLR holds—even outside exponential families—we can still construct a UMP test for one-sided alternatives. This makes MLR a **general and powerful concept** for hypothesis testing.

1. UMP Tests Depend on MLR, Not Just Exponential Form

The main requirement for UMP tests is:

$f(x;\theta 1)f(x;\theta 0)$is monotone in T(X).

for all $\theta 1 > \theta 0$.

If this holds, then:

- Data can be arranged from "least to most favorable" for the alternative.
- A simple rejection region of the form T(X)> c is optimal.
- UMP tests exist *even if the distribution is not exponential*.

Thus, exponential families are *sufficient* for UMP tests, but not *necessary*.

2. Examples of Non-Exponential Families with MLR

Below are situations where MLR holds outside the classical exponential family structure.

Example 1: Shifted Distributions (General Form)

Suppose we have a density of the form:

$f(x;\theta) = f0(x-\theta)$, a **shift family**. If f0(x) is log-concave (e.g., logistic, Laplace), the likelihood ratio often satisfies MLR in the statistic:

T(X) = sample mean or location-type statistic.T(X) = \text{sample mean or location-type statistic}.T(X) = sample mean or location-type statistic.

UMP test exists for testing increases in θ:

H0:θ ≤θ0 H1:θ>θ0.

Even though the logistic distribution is *not* an exponential family, MLR holds in many cases.

Example 2: Hypergeometric Distribution (Finite Population)

Let X = number of successes in a sample drawn **without replacement**.

X ~Hypergeometric(N,K,n)

Here:

- N= population size
- K = number of successes in population
- n = sample size

For testing:

H0:K ≤ K0 H1:K > K0, the likelihood ratio increases with X:

$\frac{f(x;K0)}{f(x;K1)}$ Thus MLR holds, and: Reject H0 if X≥c.**UMP test exists**, even though the hypergeometric distribution is *not exponential*.

Example 3: Certain Lifetime Distributions in Reliability

Many reliability models (e.g., Weibull with known shape) have likelihood ratios that are monotone in the total time on test T(X) = ∑Xi.

UMP tests exist for testing: H0:θ ≤ θ0 H1:θ > θ0, even though these models are not strictly exponential families.

3. Why This Extension Is Important

Understanding that **MLR is the key requirement** provides flexibility when dealing with applied problems:

- Quality-control sampling (hypergeometric)
- Nonparametric rank tests with MLR properties
- Lifetime and reliability models
- Distributions arising from engineering and environmental applications

This helps in designing optimal tests even outside classical textbook settings.

4. General Steps to Construct a UMP Test Under MLR (Any Distribution)

If you know the distribution has MLR in statistic T(X):

- Identify the statistic T(X).
- Confirm that likelihood ratio increases with T(X).
- Find distribution of T(X) under θ0.
- Choose cutoff cαsuch that:

1. $P_{\theta 0}(T(X) > c\alpha) = \alpha$.

Use rejection rule:

2. Reject H0 if $T(X) > c\alpha$.

This works whether or not the distribution belongs to an exponential family.

7.5 WHY UMP TESTS FAIL FOR TWO-SIDED ALTERNATIVES

Up to this point, we saw that UMP tests exist for **one-sided alternatives** in distributions with the MLR property. However, for **two-sided alternatives**, even in the simplest one-parameter exponential families, **UMP tests do NOT exist**. This failure is fundamental—not a technical difficulty. No test can maximize power in **both directions simultaneously**.

1. Structure of a Two-Sided Hypothesis

A typical two-sided test looks like: $H0:\theta = \theta 0$ vs. $H1:\theta \neq \theta 0$; This means the parameter could be: **greater** than θ0, or **less** than θ0; So the alternative hypothesis contains **two distinct directions**.

2. Conflict between both directions

To understand why UMP tests fail, imagine two different alternatives:

$\theta 1 > \theta 0$ and $\theta 2 < \theta 0$. For θ1: Large values of T(X)support H1. The MP test for this direction rejects when T(X) is **large**.

For θ2: Small values of T(X) support H1. The MP test for this direction rejects when T(X) is **small**.

Thus:

- **One NP test says reject when** T(X) **is large.**
- **The other NP test says reject when** T(X) **is small.**

These two tests **contradict each other**.

- No single rule can be best for both directions.
- This makes a UMP test impossible.

3. Example: Normal Mean Test (Known Variance)

Let:

$X1,\ldots,Xn \sim N(\mu,\sigma^2)$ with known σ^2.

Test:

$H0:\mu = \mu0$ $H1:\mu \neq \mu0$..

One-sided MP tests:

For $H1:\mu > \mu0$: reject for large $\bar{X}$.

For $H1:\mu < \mu0$: reject for small $\bar{X}$.

But two-sided alternative requires BOTH:
Reject if $\bar{X} > c1$ or $\bar{X} < c2$.

There is **no single test** that is simultaneously most powerful against:all $\mu > \mu0$, and all $\mu < \mu0$.

Even this simplest case (Normal, known variance) has **no UMP two-sided test**. This is why two-sided Z-tests are **not** UMP—they are UMPU (unbiased).

4. Example: Binomial Proportion Test

Let $X \sim Bin(n,p)$
Test:

$H0:p = p0$ $H1:p \neq p0$. If $p > p0$: reject for **large** X. If $p < p0$: reject for **small** X. A two-sided test needs *both*: $X \geq c1$ or $X \leq c2$. Any test that increases power for $p > p0$ necessarily decreases power for $p<p0$ and vice versa. Thus, no UMP test exists.

5. Summary of Reasons for Failure

UMP tests fail for two-sided alternatives because:

(a) Conflicting optimal rejection regions

MP tests for $\theta > \theta0$ and $\theta < \theta0$ are opposite in direction.

(b) No single test can dominate in both regions

A test optimal on the right tail is inferior in the left tail and vice versa.

(c) MLR only gives ordering in ONE direction

MLR cannot order data for both an increase and a decrease at the same time.

(d) Rejection region becomes two-sided

UMP theory is built on *one monotone decision rule*.
Two-sided decisions violate this structure.

6. What We use instead of UMP

Because UMP two-sided tests are impossible, we rely on:

- **UMPU tests (Uniformly Most Powerful Unbiased)**
- **Likelihood Ratio Tests (LRT)**
- **t-tests (for unknown variance)**
- **Wald and Score tests**

These have good optimality properties even though they are *not* UMP.

7.6 ALTERNATIVE OPTIMALITY CONCEPTS WHEN UMP DOES NOT EXIST

When UMP tests do not exist—especially for two-sided hypotheses or multi-parameter models—statisticians rely on **other optimality criteria** that still guarantee good performance.
These substitutes are not "best for all alternatives," but they are "best under certain fairness or generality conditions." The main alternative concepts are:

- **UMPU tests (Uniformly Most Powerful Unbiased)**
- **Likelihood Ratio Tests (LRT)**
- **Invariant Tests (Using Symmetry)**

Each of these methods solves the limitation of UMP theory in a different way.

1. Uniformly Most Powerful Unbiased (UMPU) Tests

A test is **unbiased** if:

$\beta(\theta) \geq \alpha$ for all $\theta \in H1$. This prevents tests from having low power near the null. Among all unbiased tests, if one test has the highest power everywhere in H1, it is called **UMPU**.

Most common example:

Two-sided t-test and **two-sided Z-test** are **UMPU**, not UMP.

These tests avoid favoring one side of the alternative and maintain balanced behavior.

2. Likelihood Ratio Tests (LRT)

The LRT compares how well the data fits: the null model, and the best-fitting model under the alternative.

LRT statistic:

$$\Lambda = \sup_{\theta \in H0} L(\theta)$$

$$\Lambda = \sup_{\theta \in H1} L(\theta)$$

Reject H0 if $-2\ln(\Lambda)$ is too large.

Why LRT is used:

- Very general
- Works for multi-parameter problems
- Asymptotically optimal
- Does not require MLR

Example:

Testing variance in a Normal distribution, or comparing nested regression models. Although LRT is not always UMP, it often has **good power** and is widely used.

3. Invariant Tests (Using Symmetry)

When a model has symmetry (e.g., location or scale invariance), we can construct tests that respect this structure. These tests often have **optimality within the class of invariant tests**.

Example:

Testing the mean of a Normal distribution with unknown variance.

> The **t-test** is the uniformly most powerful **invariant** test under location-scale transformations.

This is why the t-test is the standard method even though no UMP test exists.

Summary

Situation	UMP Exists?	Best Alternative
One-sided exponential family	Yes	UMP test
Two-sided alternative	No	UMPU or LRT
Unknown nuisance parameters	No	LRT / t-test
Regression models	No	Wald / Score / LRT
Symmetric (location-scale) models	No	Invariant tests

7.7 CONCLUSION

In this lesson, we deepened our understanding of how UMP tests arise, why they are powerful, and importantly—why they cannot always be constructed.While Lesson 6 focused on UMP tests for one-sided alternatives, Lesson 7 extends this framework, emphasizing **structure, conditions, generalizations, and limitations**. Here are the main ideas brought together clearly.

1. Exponential Families Provide Structure for UMP Tests

We began by revisiting the **one-parameter exponential family**, where many common distributions—Normal (with known variance), Binomial, Poisson, Exponential—are included.
These distributions share:

- A clear **sufficient statistic**
- A structured likelihood
- Often, the crucial **Monotone Likelihood Ratio (MLR)** property

This structure is what makes UMP tests for one-sided alternatives possible.

2. MLR is the Key Engine Behind UMP Tests

The MLR property ensures that: As the parameter increases, the statistic T(X) also tends to increase. This creates a **single direction of evidence**, making it possible to design tests that always reject for: $T(X)>c\alpha$. This simple threshold-based rule is the heart of UMP testing.

3. UMP Tests for One-Sided Hypotheses are powerful and simple

When MLR holds, the UMP test for: $H0:\theta \leq \theta 0$ vs $H1: \theta > \theta 0$ is straightforward to derive and implement.

Examples:

- Large sample mean in Normal distribution
- Large number of successes in Binomial distribution
- Large counts in Poisson distribution

Each is a **UMP test for detecting increases** in the parameter.

This shows how optimal tests arise naturally in simple, well-structured models.

4. UMP Tests fail for two-sided alternatives

When testing:

$H0:\theta = \theta 0$ vs $H1: \theta \neq \theta 0$, the direction of departure can be: to the **right** (large T(X)), or to the **left** (small T(X)). The most powerful test for increases requires rejecting for **large** T(X), while the most powerful test for decreases requires rejecting for **small** T(X).These conflict. Thus: No single test can maximize power for all alternatives on both sides of $\theta 0$. Even in the Normal distribution, with all its simplicity, **no UMP test exists** for two-sided problems.

5. Alternative optimality criteria save the situation

Because UMP tests fail in many common settings, we rely on alternative principles:

- **UMPU tests**: Best among unbiased tests
- **Likelihood Ratio Tests (LRT)**: Very general and widely applicable

- **Wald and Score tests**: Useful in regression and large samples
- **Invariant tests**: Optimal under symmetry (e.g., t-test)

These alternative approaches provide strong, practical solutions where UMP tests cannot be used.

- Key Takeaway 1: Exponential Families Provide Structure

Many common distributions—Normal (with known variance), Binomial, Poisson, Exponential—belong to the **one-parameter exponential family**.These families naturally provide: a **sufficient statistic** T(X),a well-structured likelihood, and often the **Monotone Likelihood Ratio (MLR)** property. This structure is critical for UMP tests.

- Key Takeaway 2: MLR Enables UMP One-Sided Tests

The **Monotone Likelihood Ratio (MLR)** property states that the likelihood ratio is monotone (increasing or decreasing) in T(X) whenever the parameter changes. MLR guarantees: A consistent direction of evidence. A threshold-based rejection rule. m Existence of a **Uniformly Most Powerful (UMP)** test for one-sided alternatives. The general UMP test form is: Reject H0 if $T(X)>c\alpha$.

- Key Takeaway 3: One-Sided UMP Tests are Common and Simple

Because MLR exists in many classical models, UMP tests for one-sided alternatives are easy to construct. Typical examples: **Normal** mean test (variance known) → reject if $\bar{X}$ is large. **Binomial** proportion test → reject if X is large. **Poisson** rate test → reject if count X is large. In these settings, UMP tests provide the **strongest possible evidence** for increases in the parameter.

- Key Takeaway 4: UMP Tests do not exist for two-sided alternatives

For hypotheses like: $H0:\theta=\theta0 vs H1:\theta\neq\theta0$, there is a **fundamental conflict**: Detecting $\theta>\theta0$ requires rejecting for **large** T(X). Detecting $\theta < \theta0$ requires rejecting for **small** T(X). No single test can maximize power in **both** directions. This impossibility holds even for simple models like the Normal distribution. Thus: **There is no UMP test for two-sided alternatives in one-parameter exponential families.**

- Key Takeaway 5: Alternative optimality concepts are needed

Since UMP tests fail in many cases, other testing principles become essential:

- **UMPU tests** (Uniformly Most Powerful Unbiased)
- **Likelihood Ratio Tests (LRT)**
- **Wald and Score Tests**
- **Invariant Tests** (e.g., the t-test)

These methods provide robust and theoretically sound solutions for two-sided, complex, and multi-parameter problems.

- Key Takeaway 6: UMP is ideal, but rare

UMP tests are theoretically elegant and highly powerful **when they exist**.
However: Their existence relies heavily on MLR,They apply mainly to one-sided tests, They rarely extend to multi-parameter or two-sided cases. Thus, UMP tests represent an **ideal benchmark**, while more flexible methods like LRT or UMPU are used in practice.

7.9 SELF-ASSESSMENT QUESTIONS

1. Define the Monotone Likelihood Ratio (MLR) property.
2. What is the general form of a UMP test under MLR?
3. What characteristics of the one-parameter exponential family help in deriving UMP tests?
4. Why do UMP tests exist for one-sided hypotheses but not for two-sided hypotheses?
5. Explain in your own words why likelihood ratios cannot be uniformly ordered for two-sided alternatives.
6. What is the difference between UMP and UMPU tests?
7. Binomial Example
   $X \sim Bin(20,p)$.
   Test: $H0:p = 0.3$ vs $H1:p > 0.3$. (a) State the UMP test.(b) If $X=8$, compute the p-value.
8. **Poisson Example**
   $X \sim Poisson(\lambda)$. Test: $H0:\lambda=4$ vs $H1:\lambda>4$. If $X=9$, find the p-value (approximate using cumulative probabilities).
9. **Normal Mean Example (Known Variance)**
   A sample of size $n = 25$ from $N(\mu,9)$ gives $\bar{X} =12$.Test: $H0:\mu = 10$ vs $H1:\mu > 10$ at $\alpha = 0.05$. (a) State the UMP test. (b) Decide whether to reject H0.

7.10 Suggested Readings

1. **Testing Statistical Hypotheses** — E. L. Lehmann & J. P. Romano
2. **Statistical Inference** — George Casella & Roger L. Berger
3. **Introduction to the Theory of Statistics** — Mood, Graybill & Boes
4. **Mathematical Statistics with Applications** — Wackerly, Mendenhall & Scheaffer
5. **Theory of Point Estimation & Testing** — Hogg, McKean & Craig

Dr. M.Vijaya Lakshmi

LESSON -8

UMP UNBIASED TESTS

OBJECTIVES

Learning Objectives (By the end of Lesson 8, students will be able to):

- Define **unbiased tests** and explain the need for unbiasedness.
- Understand why UMP tests may not exist and how **UMP Unbiased (UMPU)** tests fill this gap.
- Apply **Neyman structure** to derive UMPU tests in exponential families.
- Construct UMPU tests for two-sided hypotheses.
- Recognize similarities and differences between UMP and UMPU rejection regions.

STRUCTURE

8.1 INTRODUCTION

Why unbiasedness is needed in two-sided hypotheses. For two-sided testing: If a test puts too much weight in the upper tail, it loses power for detecting decreases. If it puts too much weight in the lower tail, it loses power for detecting increases. Unbiasedness guarantees:

- Balanced sensitivity in both directions
- Fairness across the entire alternative
- No direction of the alternative is penalized

Thus, unbiased tests form the correct class in which we can search for optimal **two-sided** tests.

UMP tests and their limitations

- A Uniformly Most Powerful (UMP) test is one that has the highest chance of rejecting a false null hypothesis compared to all other tests of the same size.
- These tests work well for one-sided alternatives in certain families of distributions.
- However, for many problems such as two-sided hypotheses, a UMP test may not exist, which creates the need for other approaches.

"unbiasedness" is imposed in hypothesis testing

- A test is called *unbiased* if it does not give an unfair advantage to the null hypothesis.
- In other words, under any alternative, the probability of rejecting the null should be at least as large as under the null itself.
- This rule helps us design tests that are reliable and meaningful, especially when a UMP test cannot be found.

Role of locally most powerful (LMP) tests for local alternatives

- Sometimes the difference between the null and alternative values is very small.
- LMP tests are designed to be the most effective in detecting such small departures from the null.
- They are useful when we want sensitivity around the null value, for example, checking if a parameter is just slightly greater than the hypothesized value.

Real-world applications of these concepts

- In **medical studies**, to detect small improvements from a new drug compared to the standard.
- In **quality control**, to notice even small deviations in a production process.
- In **economics**, to test whether a new policy has a slight but real effect.
- In **scientific experiments**, unbiased tests avoid misleading results and help ensure fair conclusions.

Before know about but UMP learn Classical Neyman-Pearson Lemma, Generalized Neyman-Pearson Lemma (GNP Lemma).

CLASSICAL NEYMAN-PEARSON LEMMA

8.1.1 Statement of the Lemma

Let $H0:\theta = \theta_0$ vs $H1:\theta = \theta_1$ be two simple hypotheses. The Neyman-Pearson Lemma states:
A test that maximizes power among all level α tests is one that rejects H_0 when the likelihood ratio
$\lambda(x) = \frac{f(x;\theta 1)}{f(x;\theta 0)}$ exceeds a certain constant k. That is, the most powerful (MP) test of level α has the critical region:
$R=\{x : \lambda(x) > k\}$

8.1.2 Assumptions and Interpretation

Assumptions:

- Both H_0 and H_1, are simple hypotheses
- Likelihood functions $f(x;\theta_0)$ f and $f(x;\theta_1)$ are well-defined
- Random variable X has a known distribution under both hypotheses

Interpretation:

- The lemma identifies the best possible test (in terms of power) for a fixed significance level α.
- The decision rule depends on comparing likelihoods under both hypotheses.
- It forms the foundation for most classical hypothesis testing procedures.

8.1.3 Concept of Most Powerful (MP) Tests

A test $\phi(x)$is most powerful of level α if for all $\psi(x)$ with $E\theta_0[\psi] \leq \alpha$, $E\theta_1[\phi] \geq E\theta 1$

That is: Among all tests that control the Type I error at α, ϕ maximizes the probability of detecting H_1 (power)

8.1.4 Critical Function and Test Construction

The test function $\phi(x)$is usually defined as:

$\phi(x)=\{1,\lambda(x) > k\gamma, \lambda(x) = k_0, \lambda(x) < k\}$

Where, k is chosen such that the test has level α:

- $E\theta_0[\phi(X)]=\alpha$
- $\gamma \in [0,1]$is used for randomization when $\lambda(x)=k$ occurs with positive probability.

8.1.5 Example: Normal Distribution with Known Variance

Problem:

Test $H_0:\mu = \mu_0$ vs $H1:\mu=\mu_1$, where $\mu_1 > \mu_0$, using $X_1,X_2,\ldots,X_n \sim$ iid $N(\mu,\sigma^2),\sigma^2$ known

Step 1: Likelihoods

$$\lambda(x) = L(\mu_1)/ L(\mu_0) = \exp\left[\frac{n(\mu 1-\mu 0)}{\sigma^2}\left(\bar{x} - \frac{\mu 1+\mu 0)}{2}\right)\right]$$

Since $\lambda(x)$ is monotonic in $\bar{x}$, we can rewrite the test:

Reject H0 if $\bar{x} > c$

Step 2: Determine Critical Value

For level α, find c such that:

$P\mu_0(\bar{x} > c)=\alpha \Rightarrow c = \mu_0 + z_\alpha$.

UMP Test:

$\phi(\bar{x})=\{1,\bar{x} > c_0,\text{otherwise}\}) = 1,$

This is the most powerful test of level for testing $\mu = \mu_0$ vs $\mu = \mu_1$.

Generalized Neyman-Pearson Lemma (GNP Lemma)

8.1.6 Motivation for Generalization: Simple vs Composite Hypotheses

The classical Neyman-Pearson Lemma applies only to simple hypotheses, i.e., both H_0 and H_1 specify a single value of the parameter. In real applications, we often face composite hypotheses, such as:

$$H_0:\theta \leq \theta 0 \text{ vs } H_1:\theta > \theta_0$$

$$H_0:\mu = 0 \text{ vs } H_1:\mu \neq 0$$

Challenge:

Multiple parameter values exist under H_0 or H_1. We need a way to construct optimal tests that work uniformly across all parameter values in one set.

8.1.7 Statement and Formulation of the GNP Lemma

Let Φ be the class of test functions $\phi(x)$ such that:

$$\underset{\theta\in\Theta 0}{Sup} E\theta[\phi(X)] \leq ;\alpha;\ \theta\in\Theta 0 \sup E\theta[\phi(X)] \leq \alpha$$

Then, the Generalized Neyman-Pearson Lemma states:

A test $\phi*\in\Phi$ is Uniformly Most Powerful (UMP) if:

$$\underset{\theta\in\Theta 1}{Inf} E\theta[\phi*(X)] \geq \underset{\theta\in\Theta 1}{Inf} E\theta[\phi(X)], \forall\phi\in\Phi$$

Key idea:

We seek a test that:

Controls the Type I error over all $\theta\in\Theta_0$. Maximizes the minimum power over all $\theta\in\Theta_1$

Formulation Summary:

Θ_0: Null parameter space (composite)

Θ_1: Alternative parameter space (composite)

ϕ: Test function

Goal: Find $\phi \in \Phi$ such that: Power $\phi = \inf_{\theta \in \Theta 1} E_\theta[\phi(X)]$ is maximized

8.1.8 Application: Finding UMP Tests in Constrained Settings

Problems where the test must perform uniformly well across a range of values. Especially useful in:One-sided tests with composite nulls, Invariance-based tests (location, scale families) and Constrained parameter problems.

Example:

Suppose $X \sim N(\mu,\sigma^2)$, $H_0: \mu \leq \mu 0$ vs $H_1: \mu > \mu_0$ Use Generalised NP Lemma and monotonicity of power in μ to justify:

$$\phi(\bar{x}) = \{1, (\bar{x} > c_0, \phi(\bar{x}) = \{1, 0, \bar{x} > c, \bar{x} \leq c\}$$

Choose 'c' to control level: $P\mu = \mu_0 (\bar{x} > c) = \alpha$

8.1.9 Limitations and Comparison with Classical Lemma

Feature	Classical NP Lemma	Generalized NP Lemma
Type of Hypotheses	Simple vs Simple	Composite (both H_0 and H_1)
Test derived via	Likelihood ratio	Infimum/supremum of expectations
Guarantees	Point wise power maximization	Uniform performance over parameter spaces
Randomization	Sometimes needed	More common
Practical Use	Direct formula	Often guides theory or asymptotics

Limitations:

- Often hard to compute infimum/supremum in practice
- UMP test may not exist even if GNP lemma gives the form
- In practice, other methods (like Likelihood Ratio Tests or invariance principles) may be more feasible

Summary

- GNP Lemma extends NP Lemma to composite hypotheses
- It helps construct UMP tests where NP Lemma cannot be directly applied
- Practical application depends heavily on the structure of parameter space
- It lays the foundation for UMP tests under monotone likelihood ratio, explored in the next lesson

8.2 UMP UNBIASED (UMPU) TESTS

Definition of Uniformly Most Powerful (UMP) Tests

- A UMP unbiased test is the **best test among all unbiased tests** of the same size.
- It is called "uniformly most powerful" because it gives the highest power for all parameter values in the alternative, but only within the class of unbiased tests.

Let $\Phi\alpha$ be the set of all level-α tests for testing $H0:\theta\in\Theta_0$ against $H1:\theta\in\Theta_1$. A test function $\phi*(x)\in\Phi\alpha$ is said to be Uniformly Most Powerful (UMP) if:

$E\theta[\phi*(X)] \geq E\theta[\phi(X)]$ for all $\theta\in\Theta 1$ and all $\phi\in\Phi\alpha$

That is, $\phi*$ has the highest power function uniformly over all values in the alternative space $\Theta 1$, among all tests that control the size at level α.

Note:

UMP tests may not always exist, especially in two-sided alternatives or multi-parameter families.

Limitations of UMP tests for two-sided hypotheses

- For one-sided alternatives (e.g., testing if mean > 0), a UMP test often exists.
- But for two-sided hypotheses (e.g., mean ≠ 0), a UMP test generally does **not** exist because no single test is best for both directions.
- In such cases, we rely on **UMP unbiased tests** to ensure fairness and optimality.

Example: Two-sided test for normal mean

- Suppose we test $H0: \mu = 0$ vs. $H1:\mu\neq 0$ with known variance.
- A simple UMP test cannot be found because we need to be powerful against both positive and negative shifts.
- By imposing the unbiasedness condition, we can construct a test based on the absolute value of the test statistic (e.g., the two-sided Z-test).
- This gives us a **UMP unbiased test** for the problem.

UMP unbiased tests extend the idea of UMP tests to situations where perfect UMP tests do not exist, especially in **two-sided testing problems**.

In hypothesis testing, we aim to design tests that have the highest probability of correctly rejecting a false null hypothesis i.e., tests with the greatest power.

Importance in Hypothesis Testing

- Optimal Performance: They offer the best possible chance of detecting false null hypotheses, across the entire alternative parameter space.
- Uniqueness: If a UMP test exists, it is often unique (up to randomization).
- Simplicity in One-Sided Tests: In many problems (especially one-sided alternatives), the UMP test has a simple and interpretation form.

- Foundation for Classical Procedures: Many standard tests (e.g., one-sided Z-test, t-test, tests in exponential families) are special cases of UMP tests.
- Benchmark Tool: They serve as a benchmark against which other tests can be evaluated for efficiency and performance.

A **Uniformly Most Powerful Unbiased (UMPU)** test is:**The most powerful test among all unbias ed tests.**

This is the best possible optimality concept when UMP tests do not exist.

Difference between UMP and UMPU

UMP Tests	UMPU Tests
Best among *all* tests of size α	Best among *all unbiased* tests of size α
Exist mainly in one-sided cases	Designed mainly for two-sided cases
Uses MLR property	Uses unbiasedness + similarity

UMP is stronger, but often impossible; UMPU is achievable.

UMPU tests typically exist in:

- One-parameter exponential families
- Two-sided hypotheses
- Problems where a complete sufficient statistic is available
- Situations where similarity conditions can be imposed
- Many classical two-sided tests (Z-test, t-test, exact Binomial test) are UMPU.

8.3 NEYMAN STRUCTURE FOR UMPU TESTS

The **Neyman Structure** provides a constructive method for deriving UMPU tests.

Conditioning on sufficient statistics

In exponential families, the sufficient statistic can be decomposed so that:

- One part contains information about the null
- The other can be conditioned on to remove nuisance parameters

This leads to conditional distributions that are easier to test.

Exact test construction

Steps:

1. Identify a complete sufficient statistic for θ.
2. Condition on that statistic to eliminate nuisance parameters.
3. Choose a rejection region RRR such that

$$P\theta 0(X \in R)=\alpha.$$

Ensure

$$\beta(\theta) \geq \alpha \text{ for all } \theta \in H1$$

to satisfy unbiasedness.

Relation to exponential families

- Exponential families have natural sufficient statistics.
- These statistics allow **exact** unbiased tests.
- UMPU tests in exponential families arise from this conditioning framework.

This is why UMPU tests are standard in classical models.

8.4 SIMILARITY AND SIMILAR REGIONS

Tests with constant size across **H0**

A test is **similar** if:

$P\theta(X \in R)=\alpha$ for all $\theta \in H0$.

This requirement ensures that:

The test maintains **exact size** across the whole null hypothesis.

No part of the null is unfairly penalized or advantaged.

Importance in two-sided testing

Two-sided alternatives require:

- Equal sensitivity to increases and decreases
- Balanced rejection region
- No bias toward one direction

Similarity ensures this balance.

Thus:

Similarity + Unbiasedness → UMPU optimality

8.5 EXAMPLES OF UMPU TESTS

1. Two-sided Normal mean test (Z-test or t-test

- Variance known → two-sided Z-test
- Variance unknown → two-sided t-test

Both are UMPU tests.

Rejection region:

$Z = \frac{\bar{X}-\mu 0}{\frac{\sigma}{\sqrt{n}}} > z\alpha/2$

or

$|t| > t_{\alpha/2,n-1}$..

2. Binomial two-sided test

Testing:

H0:p=p0 vs $p \neq p0$

The exact binomial test with equal tail areas is **UMPU**.

3. Poisson two-sided test

Testing:

H0:$\lambda = \lambda 0$ vs $\lambda \neq \lambda 0$.

An exact two-tailed Poisson test yields a UMPU test.

Why these are UMPU

All three models are:

- One-parameter exponential families
- Having complete sufficient statistics
- Allowing similarity + unbiasedness

Hence UMPU tests arise naturally from their structure.

8.6 Properties and Interpretation

Unbiasedness

The test never has lower power in the alternative than at the null:

$\beta(\theta) \geq \alpha \forall \theta \in H1$.

This guarantees fairness.

Optimality within a restricted class

The UMPU test is the **best** (most powerful) *among all unbiased tests*. Because UP tests do not exist in two-sided settings, UMPU tests are the strongest available.

Other key properties

- Two-sided rejection region
- Often symmetric around the null parameter
- Based on similar regions
- Derived using Neyman structure
- Exact (not asymptotic)

These properties make UMPU tests both practical and theoretically sound.

One-Parameter Exponential Family

A probability distribution belongs to the exponential family if its probability density function (pdf) or probability mass function (pmf) can be expressed in the form:

$f(x;\theta)=h(x)\exp[\eta(\theta)\cdot T(x)-A(\theta)]$

Components:

Θ: The natural parameter (real-valued for one-parameter case)

$T(x)$: A sufficient statistic for θ

$\eta(\theta)$: The canonical parameter, possibly a transformation of θ

$A(\theta)$: The log-partition function (ensures integration/summation equals

$h(x)$: The base measure, does not depend on θ

This form is very useful for:

- Deriving sufficient statistics
- Applying the Neyman-Fisher factorization theorem
- Establishing Monotone Likelihood Ratio (MLR) properties

Examples: Normal, Binomial, and Poisson Distributions

(a) Normal Distribution (Known Variance)

Let $X \sim N(\mu,\sigma^2)$, where σ^2 is known.

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

This can be rewritten as:

$f(x;\theta) = h(x)\exp[\eta(\theta)\cdot T(x) - A(\theta)]$

Where:

$T(x) = x$, $\eta(\mu) = \frac{\mu}{\sigma^2}$, $A(\mu) = \frac{\mu}{A\sigma^2}$ and $h(x) = \frac{e^{\frac{-1}{2\sigma^2}(x_i)^2}}{\sqrt{2\pi\sigma^2}}$

(b) Binomial Distribution

Let $X \sim Bin(n,p)$, with fixed n

$P(X = x) = (n_{c_x})\, p^x\, (1-p)^{n-x}$

Rewritten as:

$f(x;\theta) = h(x)\exp[\eta(\theta)\cdot T(x) - A(\theta)]$

Where:

$T(x) = x$

$\eta(p) = \log(\frac{p}{1-p})$

$A(p) = n\log(1+\exp(\eta(p))) = -n\log(1-p)$

$h(x) = n_{c_x}$

(c) Poisson Distribution

Let $X \sim Poisson(\lambda)$

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

Expressed in exponential family form:

$f(x;\lambda) = h(x)\exp[\eta(\lambda)T(x) - A(\lambda)]$

Where:

$T(x) = x$,

$\eta(\lambda) = \log(\lambda)$,

$A(\lambda) = \lambda$ and $h(x) = \frac{1}{x!}$

Summary

Many standard distributions fall under the one-parameter exponential family form. This structured representation is powerful and practically useful, as it helps in:

- Deriving sufficient statistics via the factorization theorem
- Identifying the Monotone Likelihood Ratio (MLR) property in the sufficient statistic
- Constructing Uniformly Most Powerful (UMP) tests, especially for one-sided alternatives

8.7 CONCLUSION

UMP tests provide ideal optimality for one-sided problems, but they completely break down for two-sided hypotheses. This failure occurs because the best rejection rule for detecting values **greater** than the null parameter uses **large** values of the test statistic, while the best rule for detecting values **less** than the null uses **small** values. A single test cannot simultaneously maximize power in **both directions**, making a UMP test impossible for two-sided alternatives.

To overcome this, statisticians restrict attention to **unbiased tests**, which guarantee that the test's power is never lower in the alternative than at the null boundary. Unbiasedness ensures **fairness** and **balanced sensitivity** toward both upward and downward departures from the null value. This balance is essential in symmetric two-sided testing situations.

Within this class of unbiased tests, we can identify a test that is **best**—the **Uniformly Most Powerful Unbiased (UMPU)** test. UMPU tests achieve maximum power among all unbiased competitors and therefore represent the strongest possible tests in settings where UMP tests cannot exist.

The construction of UMPU tests relies on two key ideas:

- **Similarity**, which ensures the test has the correct size for every value in the null hypothesis;
- **Neyman structure**, which uses conditioning on sufficient statistics to eliminate nuisance parameters and obtain exact, unbiased tests.

These principles are the foundation behind many well-known two-sided procedures. In fact, several classical tests—such as the two-sided Z-test, t-test, and the exact two-sided tests for Binomial and Poisson distributions—are all examples of **UMPU tests**.

Thus, in summary:

UMPU tests extend the concept of optimality to the two-sided setting, providing the best possible tests when UMP optimality is impossible.

8.8 SELF-ASSESSMENT QUESTIONS

1. Why can't UMP tests exist for two-sided alternatives?
2. Define an unbiased test.
3. What makes a test UMPU rather than UMP?

4. What is the Neyman structure, and how does it help develop UMPU tests?
5. Explain the concept of a "similar test."
6. Why are two-sided Z-tests and t-tests considered UMPU?
7. Give an example of a UMPU test in the Binomial or Poisson setting.
8. How do unbiasedness and similarity together ensure fair testing?

8.9 SUGGESTED READINGS

- **Testing Statistical Hypotheses** — Lehmann & Romano
- **Statistical Inference** — Casella & Berger
- **Introduction to the Theory of Statistics** — Mood, Graybill & Boes
- **Mathematical Statistics with Applications** — Wackerly, Mendenhall & Scheaffer
- **Theory of Point Estimation & Testing** — Hogg, McKean & Craig

Dr. M.Vijaya Lakshmi

LESSON -9

LMP Tests (Locally Most Powerful Tests)

OBJECTIVES

Learning Objectives (By the end of Lesson 9, students will be able to):

- Define **Locally Most Powerful (LMP)** tests.
- Understand why LMP tests are used when global UMP tests do not exist.
- Derive LMP tests using the **derivative of the power function**.
- Apply LMP methods in small-departure (local alternative) contexts.
- Explain the idea of LMP (Locally Most Powerful) tests and when they are useful.
- Describe the concept of *similar regions* in hypothesis testing.
- Compare LMP tests with UMP and UMPU tests.
- Understand Neyman's structure for constructing tests and its application.
- Solve illustrative examples involving UMP unbiased and LMP tests.

STRUCTURE

9.1 INTRODUCTION

Earlier lessons showed:

- UMP tests exist only under strict conditions (one-sided, MLR).
- UMPU tests solve two-sided problems but require unbiasedness and often conditioning.

However, in many real scenarios:

- UMP may not exist even for **one-sided** alternatives.

- Unbiasedness may be too restrictive or not required.

We may want a test that is **most powerful for alternatives that are close to the null**.This leads to the idea of **Locally Most Powerful (LMP)** tests.

LMP tests maximize power **at the null**, in the infinitesimal neighborhood of $\theta = \theta 0$.
They provide the best performance for detecting **small deviations**:

H1:$\theta=\theta 0+\delta(\delta$ small).

This is useful when:

- Alternatives are known to be slight changes
- Sample size is small
- Exact UMP tests don't exist
- The direction of the alternative is known

LMP tests are the "best available" when global UMP tests fail.

9.2 DEFINITION OF LMP TESTS

Let the **power function** be $\beta(\theta)$. At $\theta 0$, all size-α tests satisfy: $\beta(\theta 0) = \alpha$.

A test is **Locally Most Powerful (LMP)** if its derivative of the power function at the null is maximized: $\beta'(\theta 0)$ is as large as possible.

Equivalently:

The LMP test has the steepest immediate rise in power when θ moves away from $\theta 0$
in the direction of the alternative.

This is a **local** optimality criterion, unlike UMP which is **global**.

- A Locally Most Powerful (LMP) test is designed to be the **best test for detecting very small departures** from the null hypothesis.
- "Locally" means the test is most powerful in a neighborhood very close to the null value.
- It is useful when the alternative hypothesis is not far from the null, but we still want to catch small differences.

When LMP tests are useful (local alternatives near H_0)

- In real problems, sometimes the parameter may differ only slightly from the null value.
- For example, testing if a new medicine improves survival by a small margin.
- In such cases, LMP tests are preferred because they maximize the chance of detecting those **tiny shifts**.

Mathematical derivation using Taylor expansion around null

- The idea is based on expanding the **power function** of a test around the null hypothesis.
- The test that makes this slope (rate of increase in power) the largest at H_0 is called the LMP test.
- This involves using derivatives of the likelihood function, leading to tests based on the **score function**.

Solving an LMP test problem

- Problem: Test H0:$\mu=0$ vs H1:$\mu>0$ for $X\sim N(\mu,\sigma^2)$, variance known.
- Solution outline:
    1. For very small alternatives close to 0, the test with maximum slope in power function is desired.
    2. Use the test statistic $Z = \frac{\bar{X}}{\sigma/\sqrt{n}}$.
    3. Reject H0 if $Z > z\alpha$.
    4. This is the **LMP test** for local alternatives near $\mu=0$.

9.3 DERIVATION OF LMP TESTS

Let the likelihood be $L(\theta)$, and consider alternatives close to $\theta 0$:

$\Theta = \theta 0 + h$, $h > 0$, h small.

Using Taylor expansion: $L(\theta 0+h) \approx L(\theta 0) + hL'(\theta 0)$.

The most powerful test for local alternatives rejects for large values of the score.

Thus the **LMP test statistic** is:

$U(X,\theta 0)$.

Reject H0 when:

$U(X,\theta 0) > c\alpha$.

Where $c\alpha$ is chosen to ensure the test has size α.

9.4 FORM OF LMP TEST STATISTIC AND *SIMILAR REGIONS* IN HYPOTHESIS TESTING

Thus:

✓ **LMP tests are** score-type tests

They are the earliest form of what later becomes the **Score Test** or **Lagrange Multiplier Test** in advanced settings.

✓ **LMP tests reject for** large values of the score

Indicating evidence that the parameter is increasing beyond θ0\theta_0θ0.

✓ **The LMP test does not require MLR**

Identifying and using similar regions in practice

- Example: Suppose X∼ Binomial (n, p) and we test H_0:p=0.5.
- A rejection region is chosen such that P (reject H_0|p=0.5)=α.
- If this probability remains the same for all values of p in H_0, the region is **similar**.
- Such regions ensure the test keeps the correct size no matter which value in H0 is true.

Definition of similar regions in hypothesis testing

- A region is called *similar* if the probability of the test statistic falling in that region is the **same for all parameter values under the null hypothesis**.
- In other words, the test keeps the significance level constant across all cases of the null.

Intuition: maintaining significance level for all θ∈H_0

- When the null hypothesis includes more than one parameter value, we want our test to behave fairly for all of them.
- Similar regions guarantee that the test does not accidentally favor one part of the null space over another.

Role in deriving UMP unbiased tests

- In many two-sided testing problems, a UMP test does not exist.
- By restricting ourselves to *similar regions*, we can ensure the test is unbiased and has a fixed size across all values of H_0.
- This condition is key in constructing UMP unbiased tests.

Example: Similar regions in exponential families

- For distributions belonging to the **exponential family** (like normal, binomial, Poisson), it is possible to identify similar regions.
- Example: In testing H_0:μ=0 vs H_1:μ≠0 for a normal distribution, the rejection region based on |Z| is a similar region because the probability of rejection remains the same for all μ=0.

Similar regions help ensure that tests remain **fair and valid** across all null values, and they are essential in building UMP unbiased tests.

9.5 EXAMPLES OF LMP TESTS

Example: LMP test in normal distribution (mean testing)

- Suppose we test H_0:μ=0 against H1:μ>0 when variance is known.
- A one-sided Z-test is not just UMP but also LMP, since it is the most powerful test in the neighborhood of μ=0.
- If the alternative is only slightly greater than zero (say μ=0.1), the LMP test has the highest chance of detecting it.

LMP tests are **specialized tools** for alternatives very close to the null, giving us maximum sensitivity for detecting small but important differences.

Example 1: Normal Distribution (Mean Known Variance)

Let:

X∼N(μ,σ2)

Test:

H0:μ=μ0 vs H1:μ>μ0.

Compute the score: U(X, μ0) = $\frac{X-\mu 0}{\sigma^2}$.

Reject for large U(X,μ0):

X>c.

This yields the familiar one-sided Z-test:

$\frac{X-\mu 0}{\sigma}$ > zα.

Interpretation:

LMP coincides with UMP when UMP exists.

Here, since MLR holds, the LMP test = UMP test.

Example 2: Exponential Distribution

Let:

X1,...,Xn ∼Exponential

Test:

H0:θ =θ0 vs H1:θ >θ0.

The score: U(θ0) = $\frac{\partial}{\partial\theta}$ln(L(θ)) = nθ0 −∑Xi$\theta_0^2$..

Reject when the score is large:

$\sum Xi <$ constant.So the LMP test rejects for **small** sample sums.

Example 3: Binomial Distribution

Let:

X~Bin(n,p)

Test:

H0:p = p0 H1:p > p0.

Score:

$U(p0)=\frac{X-np0}{p0(1-p0)}$..

Reject when X is large.

Again, LMP test matches UMP when MLR holds.

9.6 NEYMAN'S STRUCTURE

Concept of Neyman's structure in hypothesis testing

- Neyman introduced a systematic way to construct good tests.
- His idea was to use likelihood ratios and critical regions to build tests with controlled size (significance level) and maximum power.
- This approach gives a general framework for hypothesis testing.

General framework for test construction

- Step 1: Choose the significance level α (probability of rejecting H_0 when it is true).
- Step 2: Construct a rejection region so that the probability of falling into it under H_0 is exactly α.
- Step 3: Among all such tests, pick the one that has the greatest power for detecting alternatives.

Relation to UMP unbiased and LMP tests

- UMP tests can be derived directly using Neyman's framework when they exist.
- For cases where UMP tests do not exist, adding conditions like *unbiasedness* and *similar regions* helps us find UMP unbiased tests.
- LMP tests can also be explained through Neyman's structure when alternatives are very close to H_0.

Application: Example problems using Neyman's structure

- Example: Testing the mean of a normal distribution with known variance.
  - Using Neyman's structure, we build the rejection region around the tails of the distribution to maintain level α.
  - This leads naturally to the Z-test, which is the most powerful test in this setting.
- In more complex cases, Neyman's structure guides us to construct tests that are either UMP unbiased or LMP, depending on the problem.

Neyman's structure provides a **blueprint for building optimal tests**, and it connects directly to the ideas of UMP, UMP unbiased, and LMP tests.

✓ LMP vs UMP

UMP: globally optimal

LMP: optimal only for very small deviations from θ0\theta_0θ0

When UMP exists, LMP = UMP.
When UMP does not, LMP is the next-best option.

✓ LMP vs UMPU

UMPU used for two-sided hypotheses

LMP used for one-sided, near-null testing

LMP does **not** require unbiasedness

✓ LMP vs Likelihood Ratio / Score / Wald Tests

LMP forms the foundation for:

Score test (derivative of log-likelihood)

Lagrange Multiplier test

Asymptotic tests in large samples

Thus LMP tests connect classical testing to modern, general-purpose methods.

9.7 CONCLUSION

Locally Most Powerful (LMP) tests occupy an important middle ground in statistical hypothesis testing, especially in situations where uniformly most powerful (UMP) tests do not exist. As we have seen in earlier lessons, the Neyman–Pearson Lemma provides a clear and elegant solution for constructing most powerful tests in the case of simple hypotheses. However, when hypotheses become composite—particularly one-sided composite—or when

the distribution does not have the monotone likelihood ratio (MLR) property, a UMP test often does not exist. Similarly, for two-sided alternatives, UMP tests are impossible because the directions of evidence required for detecting increases and decreases in a parameter are fundamentally incompatible. Because full, global optimality may be unattainable in these cases, statisticians often focus instead on **local** optimality—how the test performs for alternatives *infinitesimally close* to the null hypothesis. This approach leads naturally to LMP tests.

The idea behind LMP tests is straightforward yet powerful: instead of maximizing power everywhere in the alternative, we maximize the **initial slope of the power function** at the null parameter value. This is a realistic goal when we expect departures from the null to be small or when detecting early, subtle deviations is important. Formally, among all size-α tests, an LMP test ensures that the derivative of the power function at the null, denoted $\beta'(\theta_0)$, is as large as possible. In other words, if the true parameter moves away from θ_0 by a very small amount, the LMP test will be the one that reacts most rapidly and most efficiently. This approach is especially valuable in scientific disciplines where early detection of a slight change is critical, such as quality control, medical diagnostics, environmental monitoring, or reliability engineering.

A remarkable feature of LMP tests is that they are based on the **score function**, which is the derivative of the log-likelihood function with respect to the parameter. The score reflects how sensitive the likelihood is to small changes in the parameter near the null value. If the score tends to be large for values of the sample that support the alternative, then rejecting for large values of the score makes the test naturally aligned with the direction of evidence. The LMP rule therefore becomes: *Reject the null when the score function is sufficiently large*. This use of the score function immediately connects LMP tests to a major component of modern statistical inference—the **Score Test**, also known as the **Lagrange Multiplier Test**. Indeed, the score test used in regression models, generalized linear models, and many likelihood-based procedures is a direct extension of the LMP idea to multi-parameter and large-sample settings.

Another important property of LMP tests is that **they often coincide with UMP tests when UMP exists**. The same statistical structure that gives rise to UMP tests—particularly the MLR property—also ensures that the LMP test points in the correct rejection direction. Thus, the LMP test matches the UMP test in problems like testing the mean of a normal distribution with known variance or testing the parameter of an exponential distribution. In these cases, LMP tests act as a check or confirmation of the UMP result, showing that both global and local optimality criteria select the same test. This dual justification strengthens the theoretical basis for using classical tests such as the one-sided Z-test and t-test.

However, the true value of LMP tests appears in scenarios where UMP tests are not available. When the likelihood ratio is not monotone or when multiple parameters are involved, constructing a UMP test is either impossible or impractical. In such settings, the LMP test offers a meaningful and attainable optimality principle. Instead of asking, “What is the globally best test for all alternatives?”—a question that may not have an answer—we instead ask: “What is the best test for alternatives that lie very close to the null?” This local perspective avoids the contradictions that arise in two-sided or multi-parameter settings, yet still grounds the test in rigorous optimality.

LMP tests also play an important foundational role in **asymptotic theory**. In large samples, the power function can often be approximated by its first derivative at the null, meaning that local properties of the test become dominant. As a result, many widely used large-sample tests—the Wald test, the Score test, and the Likelihood Ratio test—are deeply connected to LMP ideas. Among these, the Score test most directly reflects the LMP structure, since it rejects for large values of the score statistic scaled by the observed Fisher information. Because of this, the Score test is particularly effective when testing involves nuisance parameters fixed under the null or when estimating parameters under the full model is computationally difficult.

In summary, Locally Most Powerful tests provide a crucial bridge in statistical theory between the ideal but limited world of UMP tests and the more flexible framework of UMPU tests. They are **based on the score function**, **maximize power for alternatives close to the null**, **coincide with UMP tests when UMP exists**, and **form the theoretical foundation for modern likelihood-based tests**, especially the Score test. By focusing on local optimality, LMP tests offer a practical and elegant solution in many realistic testing situations where global optimal solutions do not exist. This makes LMP testing one of the most important intermediate concepts in the theory of hypothesis testing.

9.8 SELF-ASSESSMENT QUESTIONS

1. Define a Locally Most Powerful (LMP) test.
2. Why do we need LMP tests when UMP tests do not exist?
3. What is the role of the score function in LMP tests?
4. How is an LMP test different from a UMP test?
5. Derive the score function for a Normal distribution with known variance.
6. In which situations would an LMP test be most appropriate?
7. Explain why LMP tests are “local” optimality tests.
8. How do LMP tests relate to the score (Lagrange multiplier) test?
9. Describe the concept of *similar regions* in hypothesis testing.
10. Explain Neyman’s structure for constructing tests and its application.

9.9 SUGGESTED READINGS

- **Testing Statistical Hypotheses** — Lehmann & Romano
- **Statistical Inference** — Casella & Berger
- **Introduction to the Theory of Statistics** — Mood, Graybill & Boes
- **Mathematical Statistics with Applications** — Wackerly, Mendenhall & Scheaffer
- **Theory of Point Estimation & Testing** — Hogg, McKean & Craig

Dr. M.Vijaya Lakshmi

LESSON -10

LIKELIHOOD RATIO TESTS (LRT)

OBJECTIVES

- Define the likelihood ratio test and explain its construction.
- Derive and interpret the likelihood ratio statistic for simple vs. composite hypotheses.
- Understand key properties of the LRT (consistency, invariance, optimality in certain cases).
- Explain the asymptotic distribution of the LRT statistic (chi-square approximation).
- Apply the LRT to real examples and interpret results in statistical decision-making..

STRUCTURE

10.1 INTRODUCTION

Hypothesis testing in parametric models often relies on the **likelihood function**: $L(\theta)=f(x1,x2,\ldots,xn;\theta)$, which measures how well the parameter θ explains the observed data.

Key ideas:

Larger likelihood → better fit

Likelihood maximization → Maximum Likelihood Estimator (MLE)

Hypothesis testing compares the likelihoods under restricted and unrestricted models

This naturally leads to the **Likelihood Ratio Test (LRT)**.

Motivation for Likelihood-Based Testing

- The **likelihood function** measures how well different parameter values explain the observed data.
- In hypothesis testing, we want to compare how likely the observed data are under the **null hypothesis (H_0)** versus the **alternative hypothesis (H_1)**.
- The **likelihood ratio test (LRT)** formalizes this comparison by taking the ratio of the maximum likelihood under H_0 to the maximum likelihood under H_1.
- The intuition:
  - If the data fit **much better** under H_1 than under H_0, the ratio will be **small**, leading us to reject H_0.
  - If the data fit **almost equally well** under H_0, then the ratio will be **close to 1**, and we do not reject H_0.
- The LRT provides a **general and systematic framework** for hypothesis testing, applicable across many statistical models (means, variances, regression, etc.).
- It is considered a **powerful and flexible approach**, especially when UMP tests are not available.

Distinction Between Simple and Composite Hypotheses

- **Simple Hypothesis**: A hypothesis that completely specifies the distribution of the data.
  - Example: H_0:μ=10 in a normal distribution with known variance.
  - Here, all parameters are fixed under H_0.
- **Composite Hypothesis**: A hypothesis that does not fully specify the distribution; some parameters remain unknown.
  - Example: H0:μ=10 in a normal distribution with **unknown variance**.
  - Here, variance is not specified and must be estimated from the data.
- Why this matters for LRT:
  - In **simple vs. simple** cases, we can directly compare likelihoods of fully specified models.
  - In **simple vs. composite** or **composite vs. composite**, we must **maximize** the likelihood under both H_0 and H_1 before taking the ratio.

Thus, the LRT is especially valuable in handling **composite hypotheses**, where traditional UMP tests often fail to exist.

10.2 CONSTRUCTION OF THE LRT

Likelihood Ratio Test Principle

For testing:

H0:$\theta \in \Theta 0$ vs H1:$\theta \in \Theta 1$,

define:

Likelihood Ratio Statistic

$$\lambda(x) = \frac{sup_{\theta \in \Theta 0} L(\theta)}{sup_{\theta \in \Theta} L(\theta)}$$

Interpretation:

- Numerator: best likelihood under the null
- Denominator: best likelihood under the entire parameter space

Since the numerator $\leq$ denominator, $\lambda \in [0,1]$.

LRT Rule

Reject H0 when: $-2\ln\lambda > c\alpha$,

where $c\alpha$ is chosen to obtain size α.

- $-2\ln\lambda$ has an asymptotic χ^2 distribution
- Easier to work with
- Always nonnegative

The LRT is one of the most general test procedures in statistics.

10.3 PROPERTIES OF LRT

Likelihood Ratio Tests have several appealing properties:

1. General Applicability

- LRTs can be applied to:
- Any parametric model
- Any hypothesis (simple, composite, multi-parameter)
- Problems with nuisance parameters
- Complex models (e.g., regression, GLMs)
- They do not require MLR or exponential family structure.

2. Invariance

LRTs are invariant under **reparameterization**:

If $\phi = g(\theta)$, the LRT for ϕ is the same as for θ. This is a major theoretical advantage.

3. Asymptotic Optimality

As sample size $\rightarrow \infty$:

- LRTs are asymptotically **most powerful** among all tests of size α\alpha$\alpha$.
- This property is due to a fundamental result by Wald.

4. Connection to Neyman–Pearson Lemma

For **simple vs simple** hypotheses:

LRT = UMP test.

Thus LRT generalizes the NP Lemma.

5. Basis for Many Practical Tests

Many standard tests are special cases of LRT:

- t-test
- F-test
- ANOVA
- Chi-square tests
- Tests in regression and GLMs
- Model comparison in likelihood frameworks

This makes LRT one of the most widely used tools in statistical practice.

Chi-Square Approximation of the LRT Statistic

- In practice, for large n , we approximate:

  $\lambda(x) \sim \chi^2_{\alpha,df}$

- This allows us to **set a critical region**:
  - Reject H0 if
- $\lambda(x) \sim \chi^2_{\alpha,df}$

  where $\chi^2_{\alpha,df}$ is the upper α - quantile of the chi-square distribution.

- **Example**:
  - Suppose we test $H_0: \mu=0$ vs. $H1: \mu \neq 0$ in a normal distribution with unknown variance.
  - Here, df = 1 (only one parameter, the mean, is being tested).
  - For large samples, the LRT statistic follows approximately a χ_1^2 distribution.
  - At significance level $\alpha=0.05$, the critical value is $\chi^2_{0.05,1} \approx 3.84$.

If $\lambda(x) > 3.84$, we reject H_0

10.4 EXAMPLES OF LRT IN PRACTICE

Example 1: Normal Mean (Variance Known)
$X1,\ldots,Xn \sim N(\mu,\sigma^2)$

Test:

$H0: \mu = \mu 0$ vs $H1: \mu \neq \mu 0$.

The MLE under:

H0: $\hat{\mu} = \mu 0$

Full model: $\hat{\mu} = \bar{X}$

Compute:

$$\lambda(x) = -2\ln\lambda(x) = \frac{n(\bar{X}-\mu 0)^2}{\sigma^2}$$

Reject when:

$$\left|\frac{(\bar{X}-\mu 0)}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}$$, which is the familiar **two-sided Z-test**.

Conclusion:
The Z-test is an LRT.

Example 2: Testing Population Means

- **Problem Setup**:
  Suppose X1,X2,…,Xn∼N(μ,σ²) with **known variance** σ².
  - Null hypothesis: H_0:μ = μ_0
  - Alternative: H1:μ ≠ μ_0
  -
- **Likelihoods**:
  - Under H_0, the MLE is fixed at μ = μ_0.
  - Under the full model, the MLE is $\hat{\mu} = \bar{X}$.
- **Likelihood ratio**:

$$\lambda(x) = \frac{L(\mu 0)}{L(\hat{\mu})}$$

- **LRT statistic**:

$$\lambda(x) = -2\ln\lambda(x) = \frac{n(\bar{X}-\mu 0)^2}{\sigma^2}$$

- **Distribution**:
  - For large nnn, λ (x)∼χ_1^2 .
- **Interpretation**:
  This coincides with the classical **z-test** for the mean.

Example 3: Testing Population Variances

- **Problem Setup**:
  Suppose $X_1,X_2,\ldots,X_n \sim N(\mu,\sigma^2)$ with **unknown mean**.
  - Null hypothesis: $H_0:\sigma^2=\sigma_0^2$
- Alternative: $H1:\sigma^2 \neq \sigma_0^2$
- **Likelihoods**:
  - Under H_0: MLE of mean is $\bar{X}$, variance fixed at σ_0^2.
- Under full model: MLEs are $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$
- **LRT statistic**:

$\lambda(x) = n * \ln\left(\frac{\sigma_0^2}{\hat{\sigma}^2}\right) + \frac{1}{\sigma_0^2}\sum(X_i - \bar{X})^2 - n$

- **Distribution**:
  - For large n, $\lambda(x) \sim \chi_1^2$.
- **Interpretation**
  Equivalent to the **chi-square test for variance**.

Applications in Regression and Other Models

- **Linear Regression Example**:
  Consider the model

$Y = X\beta + \epsilon, \epsilon \sim N(0,\sigma^2 I)$

Null hypothesis: H0: $\beta j = 0$ (test whether a specific predictor contributes).

  - Alternative: H1: $\beta j \neq 0$.
- **Likelihoods**:
  - Under H_0: Fit restricted regression (without predictor).
  - Under full model: Fit full regression (with predictor).
- **LRT statistic**:

$$\lambda = -2\ln\left(\frac{L\ restricted}{L\ full}\right)$$

  - For large samples, $\lambda \sim \chi^2_{df}$, where df = number of restrictions (predictors tested).
- **Interpretation**:
  - Equivalent to the **F-test** (or partial F-test) used in regression.
  - Generalizes naturally to logistic regression and other generalized linear models (GLMs).

10.5 ASYMPTOTIC DISTRIBUTION OF LRT STATISTIC

One of the fundamental results in statistics is **Wilks' Theorem**:

Wilks' Theorem

For large samples: $-2\ln\lambda \rightarrow \partial\chi_k^2$,

where
k = difference in dimensionality between $\Theta 1$ and $\Theta 0$.

This means:

- LRTs become **chi-square tests** for large nnn.
- Critical values are obtained from the $\chi 2$ distribution.
- The result does not depend on the underlying model.

Why this is powerful

- Provides a **unified method** for constructing tests.
- Allows hypothesis testing in complex models including regression.
- Applies even when the exact distribution is complicated.

10.6 USING LRT IN MULTI-PARAMETER MODELS

LRTs are especially important when multiple parameters are involved. **Nested Models**

Model M0is nested inside M1 if:

$\Theta 0 \subset \Theta 1$.

Example:
Testing whether a regression coefficient is zero.

LRT statistic: $\lambda = -2\ln\left(\frac{L\ restricted}{L\ full}\right)$

Under H0:

$-2\ln\lambda \sim \chi_k^2$

where k = number of restricted parameters.

This is widely used in:

- Logistic regression
- Poisson regression
- Survival analysis
- Mixed models
- Time-series models

10.7 LIMITATIONS OF LRT

Although LRT is powerful, it has some limitations:

1. Small-sample inaccuracy

Wilks' theorem is asymptotic.
For small samples, $-2\ln\lambda$ may not follow $\chi 2$ well.

2. Boundary issues

If parameters lie on the boundary (e.g., variance components), the chi-square approximation fails.

3. Non-regular models

Certain models (e.g., mixtures, change-point models) violate assumptions of Wilks' theorem.

4. Computational complexity

Finding supL(θ) may be difficult in:

- High-dimensional problems
- Non-convex likelihoods

Nonetheless, LRT is still the preferred method in most realistic situations.

10.8 CONCLUSION

The **Likelihood Ratio Test (LRT)** is one of the most fundamental and widely used tools in statistical hypothesis testing. It provides a unified framework for comparing how well different hypotheses explain the observed data.

1. Compares the Best Likelihoods Under Competing Hypotheses

The LRT is built on a simple yet powerful idea:

- Compute the maximum likelihood of the data under the null hypothesis (H_0).
- Compute the maximum likelihood under the alternative hypothesis (H_1).
- Compare these two values through their ratio.

If the data fits the alternative substantially better than the null, the ratio will be small, providing evidence against H_0.

2. Has Strong Theoretical Foundations

The LRT arises naturally from the likelihood principle and decision theory.
Key theoretical results supporting it include:

- **Neyman–Pearson Lemma**, which shows LRT is optimal for simple hypotheses.
- **Invariance and sufficiency principles**, which justify its general use.
- **Information-theoretic interpretations**, linking LRT to Kullback–Leibler divergence.

3. Asymptotically Optimal

As the sample size grows, the LRT enjoys several optimality properties:

- The test statistic **2 log(LRT)** converges to a **chi-square distribution** (Wilks' theorem).
- It becomes **Uniformly Most Powerful (UMP)** among a wide class of tests.
- It attains **maximal power** against local alternatives (locally asymptotically most powerful).

Thus, even when exact finite-sample properties are complex, LRT performs extremely well asymptotically.

4. Works Universally Across Models

Unlike specialized tests designed for specific distributions, the LRT:

- Applies to **any parametric model**, including non-normal, non-linear, or mixed models.
- Works for **simple vs simple**, **simple vs composite**, and **composite vs composite** hypotheses.
- Remains valid regardless of the form of the likelihood, as long as it is identifiable and regular.

This generality makes it the **default choice** in many modern statistical analyses.

5. Reduces to Many Classical Tests

Many well-known classical hypothesis tests are special cases of the LRT:

- **Z-test** for means → LRT for normal mean with known variance
- **t-test** → LRT for normal mean with unknown variance
- **F-test** → LRT for comparing nested regression models
- **Chi-square tests** → LRT for categorical data and contingency tables
- **ANOVA tests** → also derived from LRT framework

Thus, LRT unifies a large portion of classical statistics.

6. Backbone of Modern Statistical Inference

LRT is central to modern methods such as:

- Generalized Linear Models (GLMs)
- Logistic and Poisson regression
- Mixed-effects models
- Survival analysis (Cox models)
- Structural equation models
- Model selection and deviance analysis

In machine learning and econometrics, deviance, AIC, and other model comparison tools are derived directly from LRT principles.

7. Despite Limitations, LRT Remains Dominant

LRT has a few limitations:

- Performance may drop in **small samples**.
- **Boundary problems** (e.g., testing variance components) can distort the asymptotic chi-square distribution.
- Requires **maximum likelihood estimates**, which may be computationally heavy in complex models.

Even so, in large-sample settings, LRT remains:

- **The most general**

- **The most principled**
- **One of the most powerful**

methods for hypothesis testing.

10.9 SELF-ASSESSMENT QUESTIONS

1. Define the likelihood ratio statistic.
2. Why do we reject when $-2\ln\lambda$?
3. Explain Wilks' theorem.
4. What is the relationship between LRT and the Neyman–Pearson Lemma?
5. How does LRT relate to the Z-test or t-test?
6. In a multi-parameter model, how is the degrees of freedom determined for the chi-square approximation?
7. Give an example where LRT is not accurate due to small sample size.
8. Why is the LRT invariant under reparameterization?

10.10 SUGGESTED READINGS

- **Testing Statistical Hypotheses** — Lehmann & Romano
- **Statistical Inference** — Casella & Berger
- **Introduction to the Theory of Statistics** — Mood, Graybill & Boes
- **Mathematical Statistics with Applications** — Wackerly, Mendenhall & Scheaffer
- **Theory of Point Estimation & Testing** — Hogg, McKean & Craig

Dr. M.Vijaya Lakshmi

LESSON -11

NON-PARAMETRIC TESTS AND GOODNESS-OF-FIT METHODS

OBJECTIVES:

After studying this unit, you should be able to:

- To develop an overall understanding of non-parametric tests, which are distribution-free methods suitable for ordinal or non-normal data.
- To learn Wolfowitz's definition of non-parametric tests and understand why such tests do not depend on population distribution assumptions.
- To understand the concept and procedure of the Chi-square Goodness-of-Fit test, which compares observed and expected frequencies to check distributional fit.
- To learn the purpose and method of the Kolmogorov–Smirnov (K–S) Goodness-of-Fit test, this compares sample and theoretical distribution functions.

STRUCTURE

11.1. INTRODUCTION

Most of the standard methods of statistical inference are based on the familiar assumptions that the random variables have normal distributions. Then the given procedures are optimum. But for non-normal distributions the standard procedures may be far from optimum. In such cases non-parametric methods are used the non-parametric methods are concerned with the treatment of the population. Another term which is often used about the population. Another term which is often used interchangeably with "non-parametric" is "Distribution Free".

For the sake of definiteness, classified methods based on specific population assumptions may be termed as "parametric methods". A procedure will be called "Distribution free", if the statistic used has a distribution which does not depend on the distribution function of the population from which the sample is drawn.

Non-parametric tests are statistical methods that do not rely on any specific population distribution and are especially useful when data are ordinal, skewed, or when sample sizes are small. Wolfowitz defined non-parametric tests as procedures whose validity does not depend on the form of the underlying distribution, making them flexible and distribution-free. Among the commonly used non-parametric methods, the Chi-square Goodness-of-Fit test helps determine whether observed frequencies match expected

frequencies from a theoretical distribution, while the Kolmogorov–Smirnov test compares the sample cumulative distribution with a theoretical distribution using the maximum difference between them. Together, these methods provide powerful tools for analyzing real-world data where classical parametric assumptions may not hold.

11.2 NON-PARAMETRIC TESTS

Definition: Non-parametric tests are statistical tests that do not assume any specific population distribution and are used when data are ordinal, nominal, or when parametric assumptions (like normality) are violated.

Advantages:

1. **No assumption of normality**
   Example: Mann–Whitney U test can compare two groups even if data are skewed.
2. **Useful for small samples**
   Example: Sign test can be used even when sample size = 8 or 10.
3. **Works with ordinal and nominal data**
   Example: Chi-square test is used for categorical data like gender vs preference.
4. **Less affected by extreme values (outliers)**
   Example: Wilcoxon signed-rank test gives good results even when one or two values are very high.
5. **Simple to compute and interpret**
   Example: Run test for randomness is very easy to apply.

Disadvantages:

1. **Less powerful than parametric tests**
   Example: Mann–Whitney U test has lower power than t-test when data are normal.
2. **Ignores actual numerical values (uses ranking)**
   Example: Wilcoxon rank-sum test converts values to ranks, losing information.
3. **Limited advanced procedures**
   Example: No direct non-parametric equivalent for complex ANOVA models.
4. **Cannot compare means directly**
   Example: Kruskal–Wallis test compares medians/ranks, not means.
5. **Large samples needed for more accuracy**
   Example: Chi-square test requires expected cell frequency ≥ 5 for valid results.

11.3 WOLFOWIFZ DEFINITION OF NON – PARAMETER TEST:

The term non-parametric is, who used it to indicate that the population could not be specified by a finite number of parameters. Non – parametric test is concerned with the form of the population and not with any parametric values.

NOTE:

Thus the distribution free and on parametric not actually synonynomons terms. From theoretical considerations, it is convenient assume that the random variables have continuous distribution functions. Theoretically, then it is unnecessary to deal with ties.

11.4 CHI – SQUARE TEST FOR GOODNESS OF FIT:

A single random sample of size 'n' is drawn from a population with unknown cumulative distribution function F_x. We wish to test the null hypothesis.

H_0: $F_x(x) = F_0(x)$; for all x.

Where $F_0(x)$ is completely specified, against the gen alternatives.

H_1 : $F_x(x) \neq F_0(x)$; for all x.

In order to apply the chi-square – goodness of fit test provides probability basis for affecting the comparison and deciding whether the lack of agreement is too great to have occur by chance.

Assume that the 'n' observations have been grouped into k mutually exclusive categorise and is denoted by $f_i$and either e_i the observed and expected frequencies respectively for the i^{th} group i=1,2,…,k. The decision regarding fit is to be based on the deviations f_i - e_i.

Suggest by Pearson (1900) is the statistic

$$q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \qquad (1)$$

where, q is the chi - square test, Σ is the summation operator, O is observed frequency, e_i is expected frequency .

A large value of q would reflect an incompatibility between the observed and expected relative frequencies and therefore the null hypothesis on which the e_i were calculated should be rejected for q large.

The likelihood function of the sample then is

$$L(\theta_1, \theta_2, \dots, \theta_k) = \prod_{i=1}^{k} \theta_i f_i \; ; \qquad f_i = 0,1,2,\dots n$$

$$\sum_{i=1}^{k} f_i = N; \sum_{i=1}^{k} \theta_i = 1 \qquad (2)$$

The maximum likelihood estimates of the parameters in equation (2) are $\hat{\theta} = \frac{f_i}{n}$. The likelihood ratio statistic for this hypothesis is

$$T = \frac{L(\omega)}{L(n)} = \frac{L(\theta_1^0, \theta_2^0, \dots\dots\dots, \theta_k^0)}{L(\theta_1, \theta_2, \dots\dots\dots, \theta_k)} = \prod_{i=1}^{k} \left(\frac{\theta_i^0}{\theta_i} \right) f_i$$

The distribution of the quantity '-2logT' approximates the chi-square distribution. The degrees of freedom are k-1, since the restriction $\sum_{i=1}^{k} \theta_i = 1$ leaves only k-1 parameters in Ω to be estimated independently. We have here,

$$-2\log T = -2\sum_{i=1}^{t} f_i(\log\theta_1^0 - \log\frac{f_i}{n}) \quad (3)$$

Equation (3) as a test criterion for goodness of fit.

Example:1

The number of accidents per day in a city during 100 days was recorded:

No. of accidents (x)	0	1	2	3	4
Observed (O)	23	27	25	17	8

Test whether accidents follow a Poisson distribution at 5% significance.

Solution:

H_0: Data follow Poisson distribution.

Mean accidents (λ) = (0×23 + 1×27 + 2×25 + 3×17 + 4×8) / 100 = 174 / 100 = 1.74

Expected frequencies using Poisson distribution:
P(0) = 0.175 → E(0) = 17.5
P(1) = 0.304 → E(1) = 30.4
P(2) = 0.264 → E(2) = 26.4
P(3) = 0.153 → E(3) = 15.3
P(4+) = 0.104 → E(4+) = 10.4

Observed (O)	Expected (E)	O-E	(O-E)^2	(O-E)^2/E
23	17.5	5.5	30.25	1.73
27	30.4	-3.4	11.56	0.38
25	26.4	-1.4	1.96	0.07
17	15.3	1.7	2.89	0.19
8	10.4	-2.4	5.76	0.55

$\chi^2 = \Sigma((O - E)^2 / E) = 2.92$
Degrees of freedom = (5 - 1 - 1) = 3
Table value $\chi^2(0.05,3) = 7.815$
Inference:
Since 2.92 < 7.815, we accept H_0. The Poisson distribution is a good fit.

11.5 KOLMOGOROV – SMIRNOV TEST FOR GOODNESS OF FIT:

THEORM:

The statistic D_n, D_n^+ and D_n^- are completely distribution free for any continuous F_x.

PROOF: $D_n = \sup_x \left|S_n(x) - f_x(x)\right| = \max_x (D_n^+, D_n^-)$

Defining the additional order statistics $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$, we can write

$$X_{(i)} \le X_{(j)} \le X_{(i+1)}$$

The probability distribution of D_n, D_n^+ and D_n^-, therefore there are seen to depend only on the random variables $F_x(X_{(i)})$;i=1,2,3,..,n. These are the order statistics from the uniform distribution on (0,1), regardless of the original $F_x(X)$ as long as it is continuous because of the probability integer transformation. Thus D_n, D_n^+ and D_n^- have distributions which are independent of particular F_x.

A simple proof can be given by making the transformation $u = F_x(X)$ in D_n, D_n^+ and D_n^-. This will be left to the reader as an exercise. The above proof has the advantage of giving definitions of the Kolmogorov– Smirnov statistic in terms of order statistics.

In order to use the Kolmogorov statistics for inference their sampling distributions must be known since distributions are independent of F_x. we can assume without loss of generality that F_x is the uniform distribution on (0,1). The derivation of the distribution D_n is rather tedious. However, the approach below illustrates a number of properties of order statistics and is therefore included here (for an in testing alternative deviation).

11.5.1 KOLMOGOROV – SMIRNOV TWO SAMPLE TEST:

The Kolmogorov – Smirnov one sample test can be adopted to the two sample problem. In the one sample case, the Kolmogorov – Smirnov statistics compares the population CDF with the empirical distribution function. In the two sample case, the Kolmogorov – Smirnov well compare the two empirical distributions of the two samples.

For two random samples of m x's and n y's from continuous populations F_x and F_y respectively. We, define the two empirical distributions functions S_m and T_n. They are estimates of their populations CDF's and if

$H_0 : F_x(x) = F_y(x)$; for all x is true.

Then $S_m$and T_n are estimates of the common CDF. The one sample Kolmogorov – Smirnov procedure suggests that two sample Kolmogorov – Smirnov statistic should be as follows…

$$D_{m,n} = \max_x \left|S_n(x) - T_n(x)\right|$$

$$D_{m,n}^+ = \max_x \left|S_m(x) - T_n(x)\right|$$

and $$D_{m,n} = \max_x \left|T_n(x) - S_m(x)\right| = D_{n,m}^+$$

Smirnov statistic if we replace $T_n(x)$ by $T_0(x)$ and by Supremum. Also the empirical distribution form converges in probability to its population CDF uniformly; therefore we get the following results.

For the asymptotic null hypothesis

$$\lim_{m,n\to\infty} P\left[\sqrt{\frac{mn}{m+n}}, D_{m,n} \le z\right]$$

$$=-2\sum(-1)^{i-1}\exp\left[-2i^2, mn\right]$$

And $$\lim_{m.n\to 0} P\left[\sqrt{\frac{mn}{m+n}}\ D^{+}_{m,n} < Z\right] = \lim_{m,n\to\infty} P\left[\sqrt{\frac{mn}{m+n}}\ D_{m,n} \le Z\right]$$

$$=1-\exp\left[1-2Z^2\right]$$

Then test given by the statistics (1) are consistent again as the following types of alternatives

STATISTICS	ALTERNATIVES
$D_{m,n}$	$F_x(x) \neq F_y(x)$
$D^{+}_{m,n}$	$F_x(x) \geq F_y(x)$, with strict inequality for some x.
$D^{-}_{m,n}$	$F_x(x) \leq F_y(x)$, with strict inequality for some x.

The statistics are distribution free in the case of continuous F_x and F_y.

Example 1: Checking if data follows Uniform (0,1)

Data

0.10, 0.20, 0.25, 0.40, 0.70

Step 1: Sort (already sorted)
Step 2: Compute empirical CDF:

- For 0.10 → 1/5 = 0.20
- For 0.20 → 2/5 = 0.40
- For 0.25 → 3/5 = 0.60
- For 0.40 → 4/5 = 0.80
- For 0.70 → 5/5 = 1.00

Step 3: Compare with theoretical CDF of Uniform(0,1):
F(x) = x

Compute $|F(x) - S_n(x)|$ at each point.

Step 4: Largest difference = **D**

Step 5: Compare with K–S critical value
If **D > D(critical)** → Reject H_0
If **D ≤ D(critical)** → Accept H_0

Example 2: Testing Normality

Data:

12, 15, 18, 19, 20, 22, 24, 25 (n = 8)

H_0: Data comes from a normal distribution.
Process:

1. Standardize each value:

$$Z = \frac{x - \bar{x}}{s}$$

2. Compute theoretical CDF Φ(z)
3. Compute empirical CDF
4. Find largest vertical difference
5. Compare D with K–S table value

If D < critical value → Normal distribution is acceptable

Advantages of K–S Test

1. Distribution-free (non-parametric)

Does not depend on the form of distribution of the data.

2. No need for histogram or grouping

Uses raw data directly.

3. Works for any continuous distribution

Normal, Exponential, Weibull, Uniform, Logistic, etc.

4. Easy to compute

Only requires CDF comparisons.

5. Very useful for small sample sizes

Even n = 5, 6, 7 works.

Disadvantages of K–S Test

1. Only for continuous distributions

Not suitable for discrete data (use Chi-square instead).

2. Less powerful than other tests for normality

Shapiro–Wilk and Anderson–Darling are stronger.

3. Sensitive in the middle, not tails

It focuses on maximum vertical difference → Tail behavior is not well detected.

4. Parameters must be known

If mean and variance are *estimated* from data → K–S becomes less accurate (Lilliefors correction needed).

5. Not suitable for large samples

Even small deviations cause rejection.

11.6 CONCLUSION

In conclusion, non-parametric tests play an important role in statistical analysis when the assumptions of parametric tests—such as normality or equal variances—are not satisfied. Wolfowitz's definition highlights that these tests are distribution-free and rely mainly on ranks, counts, or order of data rather than actual numerical values. The Chi-square Goodness-of-Fit test checks how well observed frequencies match theoretical expectations, making it valuable for categorical data. Similarly, the Kolmogorov–Smirnov test evaluates the agreement between a sample distribution and a theoretical distribution using cumulative frequencies. Together, these methods provide flexible, robust, and widely applicable tools for analyzing real-world datasets that do not follow classical distributional assumptions.

11.7 SELF ASSESSMENT QUESTIONS

1. Explain the assumptions and uses of non-parametric tests.
2. Describe Wolfowitz's definition and importance of non-parametric methods.
3. Explain the procedure for the Chi-square Goodness-of-Fit test with an example.
4. Discuss the steps involved in performing the Kolmogorov–Smirnov test.
5. Compare Chi-square and Kolmogorov–Smirnov Goodness-of-Fit tests.
6. Describe the Kolmogorov–Smirnov test with assumptions, procedure, test statistic, and interpretation.
7. Compare parametric and non-parametric tests with examples and discuss when each type is appropriate.

11.8 SUGGESTED READING BOOKS:

1. Statistical Inference by H.C, Saxena & Surendran
2. An outline of Statistical theory vol.2 by A.M. Goon and B. Das Gupta.
3. An Introduction to probability and Mathematical statistics by V.K. Rohatgi.
4. Mathematical Statistics- Parimal Mukopadhyay(1996), New central Book Agency (P)Ltd., Calcutra.

Dr. Syed Jilani

LESSON -12

KENDALL'S TAU, KRUSKAL–WALLIS & FRIEDMAN TESTS

OBJECTIVES:

After studying this unit, you should be able to:

- To understand the basic concepts of non-parametric statistical methods used when data do not satisfy normality assumptions.
- To learn how to measure strength and direction of association using Kendall's Tau based on concordant and discordant pairs.
- To acquire knowledge of how rank-based tests such as Kruskal–Wallis and Friedman's ANOVA help compare multiple groups.
- To understand the purpose of the Kruskal–Wallis test in comparing three or more independent samples using median ranks.
- To understand the purpose of Friedman's test in comparing three or more related or repeated-measure samples.

STRUCTURE

12.1. INTRODUCTION

Kendall's Tau (τ) is a non-parametric measure of correlation that evaluates the strength and direction of association between two variables measured on ordinal, interval, or ratio scales.

It is based on comparing the number of concordant and discordant pairs rather than actual numerical distances. Because it uses ranks, Kendall's Tau is robust to outliers, suitable for small sample sizes, and ideal when the underlying relationship is monotonic but not necessarily linear. It is commonly used in behavioral sciences, social sciences, and quality control where data may not follow normality.

The Kruskal–Wallis H test is a non-parametric alternative to one-way ANOVA, used when comparing three or more independent groups on a continuous or ordinal outcome. It is applied when normality or homogeneity of variance assumptions are violated. The test works by ranking all observations across groups and evaluating whether the median ranks differ significantly.

It is widely used in medical research, education, psychology, and other fields where data are

not normally distributed. If the Kruskal–Wallis test is significant, it suggests that at least one group median differs**,** requiring post-hoc pairwise comparisons.

Friedman's test is a non-parametric alternative to repeated-measures ANOVA**,** used for comparing three or more related or matched groups**.** It is applied when the same subjects are measured under different conditions or times**,** and normality assumptions are not satisfied.

The method converts raw data into ranks within each block (subject) and tests whether the treatments produce significantly different median ranks**.** Friedman's test is widely used in psychology, medicine, engineering, and preference studies where repeated measures are common.

12.2 KENDALL'S TAU COEFFICIENT

The Kendall's tau a measure of association between random variables from any bivariate population and is defined as

$$\tau = \prod_e - \prod_d$$

Where, for any two independent pairs of observations (X_i,Y_i), (X_j,Y_j) from the population.

$$\prod_d = P\left[(X_i - X_j)(Y_i - Y_j) < 0\right]$$
$$\prod_e = P\left[(X_i - X_j)(Y_i - Y_j) > 0\right]$$

In order the estimate the parameter …… from a random sample of n pairs (X_1,Y_1), (X_2,Y_2),…(X_n,Y_n) drawn from this bivariate population we must find point estimates of the probabilities π_e, π_d. For each set of pairs (X_i,Y_i), (X_j,Y_j) of sample observation define the indicator variables .

$$A_{ij} = Sgn(X_j - X_i)\, Sgn(Y_j - Y_i)$$

where, Sgn(u) = - 1 if u< 0

> 0 if u = 0

= 1 if u = 0

Then the values assumed by A_{ij} are

$$\mathbf{a_{ij}} = \begin{cases} 1 \text{ if these pairs are concordent} \\ -1 \text{ if these pairs are discordent} \\ 0 \text{ if these pairs are either concordent nor discordent because of a tie in either component} \end{cases}$$

The marginal probability distribution of these indicator variables is

$$A_{ij}(a_{ij}) = \begin{cases} \prod_e & \text{if } a_{iij} = 1 \\ \prod_d & \text{if } a_{ij} = -1 \\ 1 - \prod_e - \prod_d & \text{if } a_{ij} = 0 \end{cases}$$

And then expected value is

$$E(A_{ij}) = 1(\prod_e) + (-1)\prod_d = \prod_e - \prod_d = \tau$$

Since, obviously we have $a_{ij} \leq a_{ji}$ and $a_{ij} = 0$. There are only (n_2) sets of pairs which need to be considered. An unbiased estimator of …… is therefore provided by

$$\tau = \sum_{1<i} \sum_{<j<n0} A_{ij} = \sum\sum\sum \frac{A_{ij}}{n(n-1)}$$

This means of the association in the paired observation is called Kendall's sample au coefficient.

$$Mean: E(A_{iij}) = \tau$$

$$\text{var}\, iance: v(A_{ij}) = (\Pi_e - \Pi_d) - (\Pi_e - \Pi_d)^2$$

Advantages of Kendall's Tau

1. Suitable for ordinal data

Does not require numerical/interval data.

2. More robust to outliers

Since it uses order (rank) instead of raw values.

3. Easy to understand

Based on concordance and discordance → intuitive.

4. Works well for small sample sizes

Better than Spearman's rho when **n < 30**.

5. Measures monotonic relationship

Detects increasing or decreasing patterns.

6. Distribution-free (Non-parametric)

No assumption about normality.

7. Handles ties better (Tau-b and Tau-c)

Useful in real survey or social science data.

Disadvantages of Kendall's Tau

1. More time-consuming to compute

Requires pairwise comparisons:
$\frac{n(n-1)}{2}$ pairs → slow for large datasets.

2. Less powerful with large datasets

Spearman's rho is more popular for large n.

3. Cannot handle nominal data

Needs rankings/order.

4. Lower numerical magnitude than Pearson/Spearman

Values stay smaller, sometimes misinterpreted as "weak correlation".

5. Sensitive to many ties in the data

If many tied ranks → Tau value shrinks.

6. Not suitable for non-monotonic relationships

Only detects monotonic trends.

12.3 KRUSKAL – WALLLIS TEST:

The median test for k – samples used information about the magnitude of each of the N observations relative to a single number which is the median of the pooled samples. Most of the other k – sample t_i use more of the available information by considering the relative magnitude of each observation when compare with every other observation. The comparison is offer in terms of ranks.

Since under H_0 we have essentially a single sample of size N from the common population combine the N observations into a single ordered sequence from smallest to largest, keeping track of which observation is from W sample and assign the rank 1, 2,..., N to the sequence. If adjacent – ranks are well distributed among the k sample which would be true for a random sample from a single population, the total sum of ranks $\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$ will be divided proportionally, the i^{th} sample which contains n_i observations, the expected sim of ranks would be

$$\frac{n_i}{N}\frac{N(N+1)}{2} = \frac{n_i(N+1)}{2}$$

Equivalently since the expect rank for any observation average rank $\frac{(N+1)}{2}$ for n_i observations the expected sum of ranks is $n_i(N+1)$ denoted by R_i. The actual sum of ranks assigned to the elements in the i^{th} sample. A reasonable test statistic could be based on a function of the

deviations between these observed and expected rank sums. Since deviations in either direction indicate disparity between the samples and absolute values are not particular treatable mathematically the sum of squares of theses deviations can be employed as

$$S = \sum_{i=1}^{k} \left[R_i - \frac{n_i(N+1)}{2} \right]^2 \tag{1}$$

The null hypothesis is rejected for large values of S. (In order to determine the null probability distribution of S consider the ranked sample data recorded in a table with K columns, where he entire in the i^{th} column are the n_i ranks occupied by the elements in the i^{th} sample. R_i is then the i^{th}column sum. Under H_0 the integers 1,2,3,..,N are assigned at random to the K column except for the restriction that there be n_i integers in column i). The total number of ways to make the assignment of ranks then the number of partitions of N distinct elements into K ordered sets, the i^{th} of size n_i and this is

$$\frac{N_i}{\prod_{i=1}^{k} n!}$$

Each of these probabilities must be enumerated and the values of S valuated for each. If t(S) denotes the number of assignments with the particular values 'S' calculated from equation (1) then, i=1,2,….,n $\prod_{i=1}^{n} n!$ Obviously, the calculation required are expected tedious and therefore will not be illustrated here of exact probabilities for 'S' are available in K_{ij} (1952), for k=3,4 and 5 but only for n_i equal and very critical values for some large equal sample size are also given.

Somewhat more useful test criterion is weighted sum of squares of deviation with the reciprocal of the respective sample sizes used as weights, thus the Kruskal – Wallis statistics is denoted as

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left[R_i - \frac{n_i(N+1)}{2} \right]^2 \tag{2}$$

The consistency of H is investigated Kruskal. H and S are equivalent test criteria only for all n_i equal exact probabilities for H have been tabulated in Kruskal Wallis (1952) for k = 3 all <5 and more extensive tables for k = 3 are given in Alexander and Quacle (1968).

Under the null hypothesis, the n_i entries in column were randomly selected from the set {1,2,3,…,N}. They actually constitute a random sample of size n_i drawn without replacement from the finite population consisting of the first N integers.

The mean and variance of this population are

$$\mu = \sum_{i=1}^{N} \frac{1}{N} = \frac{N+1}{2}, \sigma^2 = \sum_{i=1}^{N} \left(i - \left(\frac{N+1}{2} \right) \right)^2 = \frac{N^2-1}{12}$$

The only assumption made initially was that the population was continuous exact this is of course was it avoid the problem of ties when two or more observation are tied with in a

column, the value of H is the same regardless of the method used to resolve the ties since rank sum is not affected when tis occur across columns, the midrank method is generally used. Alternative for a conservative test the ties can be broken in the way which is least conductive to rejection of H_0

Example

We want to test whether three fertilizers give the same yield.

Data:

- **F1:** 3, 4, 5
- **F2:** 2, 3, 4
- **F3:** 6, 7, 8

Step 1: Combine and rank all values

Value	Rank
2	1
3	2.5
3	2.5
4	4
4	4
5	6
6	7
7	8
8	9

Step 2: Sum of ranks

- **F1:** 4 + 6 + 7 = **17**
- **F2:** 1 + 2.5 + 4 = **7.5**
- **F3:** 8 + 9 + 6 = **23**
  (You may recompute exactly—values may slightly vary depending on ties; idea remains same.)

Step 3: Calculate H

N=9, k=3

Compute simplified:

$$H = \frac{12}{9(10)}\left(\frac{17^2}{3} + \frac{7.5^2}{3} + \frac{23^2}{3}\right) - 3(10)$$

$H \approx 6.5$

Interpretation:

- df = k – 1 = 2
- Chi-square critical @ 5% = 5.99

Since **6.5 > 5.99**,
Reject H_0 → Fertilizer types show significant difference in median yield**.**

Advantages of Kruskal–Wallis Test

1. Non-parametric (no normality required)

Works well for skewed data, outliers, ordinal data.

2. Can compare 3 or more groups

Unlike Wilcoxon/Mann–Whitney (only 2 groups).

3. Simple to compute and interpret

Based on ranks → easy to explain and teach.

4. Robust to outliers

Ranks reduce the influence of extreme values.

5. Works with small sample sizes

Does not need large n for validity.

6. Useful when variances are unequal

Unlike classical ANOVA, equal variance is not mandatory.

Disadvantages of Kruskal–Wallis Test

1. Does not indicate *which* groups differ

Post-hoc tests (Dunn, Conover) required.

2. Less powerful than ANOVA

If data are normally distributed → ANOVA is better.

3. Requires comparable distributions

If group shapes differ (skewness), results may be misleading.

4. Cannot be used for repeated measures

Uses **independent** samples only.

5. Information loss

Converting data to ranks reduces numerical precision.

6. Sensitive when many ties occur

Ties reduce accuracy → correction needed.

12.4 FRIEDMAN'S TWO-WAY ANOVA BY RANKS:

Let us denote the ranked observations by $R_{ij}\,; i = 1,2,3,\ldots,k, j = 1,2,3,\ldots,n$. So, that R_{ij} is the rank of treatment number 'j' when observed in block number i. Then $R_{i1}, R_{i2}, \ldots, R_{kj}$ is the set of ranks given to treatment number 'j' ibn all blocks. We represent the data in tabular from as follows

TREATMENTS

	1	2	3	$\cdots$	j	$\cdots$	n	ROWS TOTAL
BLOCKS	1	R_{11}	R_{12}	R_{13}	$\cdots$	R_{1j}	$\cdots$	R_{1n}	
	2	R_{21}	R_{22}	R_{23}	$\cdots$	R_{2j}	$\cdots$	R_{2n}	
	3	R_{31}	R_{32}	R_{33}	$\cdots$	R_{3j}	$\cdots$	R_{3n}	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	
	i	R_{i1}	R_{i2}	R_{i3}	$\cdots$	R_{ij}	$\cdots$	R_{in}	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	
	k	R_{k1}	R_{k2}	R_{k3}	$\cdots$	$\cdots$	$\cdots$	R_{kn}	$\frac{n(n+1)}{2}$
COLUMN TOTAL		R_1	R_2	R_3	$\cdots$	$\cdots$	$\cdots$	R_n	$\frac{kn(n+1)}{2} \rightarrow (1)$

In the treatments area all the same, each expected colum total is the same and equals the average column total $\frac{k(n+1)}{2}$. Therefore, the sampling distribution of the random variable is the average ranks sum for the i^{th} column, R_i-R_i/n the mean of this random sample and as for any mean from a finite population.

$$E\left(\overline{R_i}\right) = \mu \;, \quad Var\left(\overline{R_i}\right) = \frac{\sigma^2 (N - n_i)}{n_i (N-1)}$$

then we have

$$E(\overline{R_i}) = \frac{N+1}{2} \;, \; V(\overline{R_i}) = \frac{(N+1)(N-n_i)}{12 n_i}$$

Since, $\overrightarrow{R_i}$ is a sample mean, if n_i is a large, the central limit theorem allows us to approximate the distribution of

$$Z_i = \frac{\overline{R_i} - \left(\frac{N+1}{2}\right)}{\sqrt{\frac{(N+1)(N-n_i)}{12n_i}}} \tag{2}$$

By the standard normal consequently z_i^2 is distributed approximately as chi-square with one degrees of freedom. This holds for i=1,2,3…,k but the z_i are clearly not independent random variables

since, $\sum_{i=1}^{k} n_i \overrightarrow{R_i} = \frac{N(N+1)}{2}$ a constant. Thus it should to be surprising that if no. n_i is very small, the random variable.

$$\sum_{i=1}^{k} \frac{N-n_i}{N} z_i^2 = \sum_{i=1}^{k} \frac{12n_i\left(\overline{R_i} - \frac{(N+1)}{2}\right)^2}{N(N+1)} = H \tag{3}$$

is distributed approximately as chi-square with (k-1) degrees of freedom. The statistic H is easier to calculate in the following from, which is algebraically equivalent to equation (2) and equation (3)

$$\text{H} = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

The rejection region is $H \geq \chi^2_{\alpha,k-1}$ some other approximations to the null distribution of H are discussed in Alexander and Quade (1968)

$$s = \sum_{j=1}^{n}\left[R_j - \frac{k(n-1)}{2}\right]^2 \sum_{j=1}^{n}\left[\sum_{i=1}^{k} R_{ij} - \frac{(n-1)}{2}\right]^2 \tag{4}$$

The probability distribution of S is

$$f_s(s) = \frac{Nf_s(s)}{(n!)^k}$$

Where 'us' is the number of these assignments which yields as the sum of squares of column total deviations. Using the symbol μ to denote $\frac{n+1}{2}$. Equation (4) can be written as,

$$S = \sum_{j=1}^{n}\sum_{i=1}^{k}\left(R_{ij} - \mu\right)^2 + 2\sum_{j=1}^{n}\sum_{1\leq i\leq p\leq k}\left(R_{ij} - \mu\right)\left(R_{ij} - \mu\right)$$

$$= k\sum_{j=1}^{n}\left(R_j - \mu\right)^2 + 2U$$

$$= \frac{kn(n^2-1)}{12} + 2U \qquad (5)$$

By spearman's Rank Correlation.

$$E(R_{ij}) = \frac{n+1}{2}, V(R_{ij}) = \frac{n^{2-1}}{12}, COV(R_{ij}, R_{iq}) = \frac{-n+1}{12}$$

Furthermore, by the design assumptions, observations in different rows are independent, so that $\forall\, i \neq p$ the expected value of a product of functions of R_{ij} and R_{pq} is the product of expected values and $Cov(R_{ij}, R_{iq}) = 0$.

Then $E(U) = n\binom{k}{2}, COV\left(R_{ij}, R_{pj}\right) = 0$

So, that $V(U) = E(U^2)$ where

$$U^2 = \sum_{j=1}^{n}\sum_{1\le i\le}\sum_{p\le k}\left(R_{ij}-\mu\right)^2\left(R_{pj}-\mu\right)^2 + 2\sum_{1\le j}\sum_{\le q\le n}\sum_{1\le i}\sum_{\le p\le k}\sum_{1\le r}\sum_{\le s\le k}\left\{\left(R_{ij}-\mu\right)\left(R_{pj}-\mu\right)\left(R_{rq}-\mu\right)\left(R_{sq}-\mu\right)\right\} \rightarrow (4)$$

Since, R_{ij} and R_{pq} are independent whenever $i \neq p$, we have

$$E(U^2) = \sum\sum\sum V\left(R_{ij}\right)V\left(R_{pi}\right) + 2\sum\sum\binom{k}{2}COV\left(R_{ij}, R_{iq}\right)COV\left(R_q\right)$$

$$= n\binom{k}{2}\frac{\left(n^2-1\right)^2}{144} + 2\binom{n}{2}\binom{k}{2}\frac{\left(n+1\right)^2}{144}$$

$$= n^2\binom{k}{2}\left(n+1\right)^2\frac{\left(n-1\right)}{144}$$

Using these results back in equation (3), we find

$$E(S) = \frac{kn\left(n^2-1\right)}{12}, V(S) = \frac{n^2k\left(k-1\right)\left(n-1\right)\left(n+1\right)^2}{12}$$

A linear function of the random variable defined as

$$F = \frac{125}{kn\left(n+1\right)} = \frac{125\sum_{j=1}^{n}R_j^{\,2}}{kn(n+1)} - 3k\left(n+1\right)$$

Has moments $E(F)=n-1, V(F)=\frac{2(n-1)(k-1)}{k}\approx 2(n-1)$, which are the first two moments of a chi-square distribution with (n-1) degrees of freedom.

The total sum of squares of deviations of all nk values around the average rank is

$$S_t=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(r_{ij}-\bar{r}\right)^2=k\sum_{j=1}^{n}\left(j-\frac{n+1}{2}\right)^2=kn\left(\frac{n^2-1}{k}\right)$$

And thus we could write Friedman's test statistic in equation (5) as

$$F=\left(\frac{n-1}{S_t}\right)S$$

Even though S_t is a constant, as in classic ANOVA problem it can be portioned into a sum of square of deviations between columns plus a error sum of squares.

The grand mean and column mean by

$$\bar{r}=\sum\sum\frac{r_{ij}}{nk}=\frac{n+1}{2}, \bar{r}_j=\frac{r_j}{k}=\sum\frac{r_{ij}}{k}$$

We have,

$$S_t=\sum\sum\left(r_{ij}-\bar{r}\right)^2=\sum\sum\left(r_{ij}-\bar{r}_{j-1}.\bar{r}_j-\bar{r}\right)^2$$

$$S_t=\sum\sum\left(r_{ij}-\bar{r}_j\right)^2+k\sum\left(\bar{r}_j-\bar{r}\right)^2+2\sum\left(\bar{r}_j-\bar{r}\right)\sum\left(r_{ij}-r_j\right)$$

$$S_t=\sum\sum\left(r_{ij}-\bar{r}_j\right)^2+\frac{\sum\left\{r_j.k\frac{(n+1)}{2}\right\}^2}{k}$$

$$S_t=\sum\sum\left(r_{ij}-\bar{r}_j\right)^2+\frac{S}{K}$$

$$S_t=kn\frac{n^2-1}{k}$$

ANOVA TABLE:

S.V	D.F	SS	MSS	F-STATISTIC
Between columns	(n-1)	s/k	s/k(n-1)	(k-1)s/k st-s
Between Rows	(k-1)	0	0	0
Errors	(n-1)(k-1)	St-S/k	St-S/k(n-1)(k-1)	
Total	nk-1	St		

There is no variation between rows here since row sums are all equal.

12.5 CONCLUSION

In conclusion, Kendall's Tau, the Kruskal–Wallis test, and Friedman's two-way ANOVA by ranks together provide a powerful set of non-parametric statistical tools for analyzing data that do not meet the assumptions of normality or equal variances. Kendall's Tau measures the strength and direction of association between two ranked variables, while the Kruskal–Wallis test compares three or more independent groups based on their median ranks. Friedman's test extends this approach to three or more related or repeated-measure samples**,** identifying differences in treatments across the same subjects.

Collectively, these methods rely on ranking rather than actual values, making them robust to outliers, suitable for ordinal data, and especially useful in real-world situations where classical parametric assumptions fail.

12.6 SELF ASSESSMENT QUESTIONS

1. Discuss Kendall's Tau coefficient in detail with its computation steps, formula, interpretation, and applications.
2. Given a set of rank data, compute Kendall's Tau and interpret the result.
3. Explain the advantages, disadvantages, and real-life applications of Kendall's Tau with examples.
4. Explain the Kruskal–Wallis test in detail, including assumptions, procedure, formula, test statistic distribution, and interpretation.
5. Using a numerical example, perform the Kruskal–Wallis test and draw conclusions.

6. Describe Friedman's test in detail, including assumptions, ranking procedure, test statistic, and interpretation.
7. Solve a numerical example using Friedman's two-way ANOVA by ranks.

12.7 SUGGESTED READING BOOKS:

1. Statistical Inference by H.C, Saxena & Surendran
2. An outline of Statistical theory vol.2 by A.M. Goon and B. Das Gupta.
3. An Introduction to probability and Mathematical statistics by V.K. Rohatgi.
4. Mathematical Statistics- Parimal Mukopadhyay(1996), New central Book Agency (P)Ltd., Calcutra.

Dr. Syed Jilani

LESSON -13

STATISTICAL METHODS FOR MODEL VALIDATION AND LARGE SAMPLE INFERENCE

OBJECTIVES:

After studying this unit, you should be able to:

- To understand the concept and purpose of Bartlett's Test**,** which is used to determine whether several independent samples have equal population variances, an important assumption for many parametric statistical procedures.
- To learn the Chi-square Test for Homogeneity of Correlation Coefficients, which examines whether correlation coefficients obtained from different independent samples represent the same underlying population correlation.
- To understand the F-Test for Linearity of Regression**,** which evaluates whether a linear regression model is adequate or whether significant lack-of-fit indicates the need for a nonlinear relationship.
- To learn the concept and application of Variance-Stabilizing Transformations**,** which are used to make the variance approximately constant across different levels of a variable, thereby satisfying assumptions of parametric tests.
- To acquire knowledge about Tests of Significance for Large Samples**,** which rely on the Central Limit Theorem and involve Z-tests for means, proportions, and variances when the sample size is sufficiently large.

STRUCTURE

13.1. INTRODUCTION

In advanced statistical analysis, several specialized tests are used to verify assumptions, compare estimates, and ensure the validity of inferential procedures. Sections 13.2 to 13.6 focus on important methods that help assess variability, correlation, regression behavior, and sample-based inference.

Bartlett's Test is used to check whether multiple independent estimates of variance come from populations with the same true variance. It is especially important when applying parametric methods like ANOVA, which assume homogeneity of variances.

The **Chi-square Test for Homogeneity of Correlation Coefficients** allows us to test whether correlation coefficients computed from different independent samples represent the same underlying population correlation. This becomes crucial in meta-analysis, reliability studies, and comparative research.

The **F-test for Linearity of Regression** is used to verify whether a linear regression model is appropriate, or whether there exists significant curvature that suggests a nonlinear relationship. It checks the adequacy of the linear form before proceeding with prediction or interpretation.

Variance-Stabilizing Transformations are mathematical transformations applied to data to make the variance approximately constant across different levels of the variable. This allows parametric tests, which assume constant variance, to be more valid and effective.

Finally, **Tests of Significance for Large Samples** rely on the Central Limit Theorem, which ensures that for large sample sizes, many statistics approximate a normal distribution. This allows the use of Z-tests and Chi-square tests for making inferences about population means, proportions, and variances.

13.2 BARTLETT'S TEST FOR HOMOGENEITY OF SEVERAL INDEPENDENT ESTIMTES OF THE SAME POPULATION VARIANCE: -

We know that the sample variance is

$$S_i^2 = \frac{1}{ni-1}\sum_{j=1}^{ni}\left(X_{ij} - \overline{X_i}\right)^2, (i = 1,2,.....,k)$$

be the unbiased estimate of the population variance obtained from the i^{th}samplex_{ij}(j=1,2,…n_i) and based on v_i=(n_i-1) d.f all the k samples being independent.

Under the null hypothesis that the samples come from the same population with variance σ^2,i.e. the independent estimates S_i^2(i=1,2,…k) of σ^2 are homogenous Bartlett's proved that the statistic

$$\chi^2 = \sum_{i=1}^{k}\left(V_i.\log_e\frac{S^2}{S_i^2}\right)\left[1-1\frac{1}{3(k-1)}\left\{\sum_i\left(\frac{1}{V_i}\right)-\frac{1}{V}\right\}\right]$$

Where,

$$S^2 = \frac{\sum V_i S_i^2}{\sum V_i}$$

Follows chi-square distribution with (k-1) DF

S^2defined in k is also an unbiased estimate σ^2,since

$$E\left(S^2\right)=\frac{\sum VE(_iS_i^2)}{\sum V_i}=\frac{\left(\sum V_i\right)\sigma^2}{\sum V_i}=\sigma^2$$

Let S_i^2 and S_j^2; $i\neq j$, $1\leq(i,j)\leq k$ be the smallest and the largest values of the unbiased estimates of σ^2 respectively. If on the basic of F-test these do not differ significantly, then all the estimates S_t^2 which lie between S_i^2 and S_j^2 won't differ significantly either and consequently all the estimates can be reasonable regarded as homogenous, coming from the same population. In this case, therefore, there is no need to apply Bartlett's test.

13.3. CHI-SQUARE TEST FOR HOMOGENEITY OF CORRELATION COEFFICIENTS:

Let $r_1,r_2,\ldots r_k$ be correlation coefficients from independent samples of sizes $n_1,n_2,\ldots n_k$ respectively we want to test the hypothesis that these sample correlation coefficients are the estimates of the same correlation coefficients from bivariate normal population obtain the value of $z_1,z_2,\ldots,z_t$
From the table of fisher's z-transformation of form

$$Z_i=\frac{1}{2}\log_e\left(\frac{1+r_i}{1-nr_i}\right)=\tanh^{-1}r_i;i=1,2,.....,k$$

These z_i's are normally distributed about a common

$$\xi=\frac{1}{2}\log_e\left(\frac{1+p}{1-p}\right)\text{ and variance}=\frac{1}{ni-3}$$

The minimum variance estimate z of the common mean of z_i's is obtained by weighting the values z_i's inversely with their respectively variances. The estimates of z is,

therefore $\overline{Z}=\dfrac{\sum\limits_i Z_i\left(ni-3\right)}{\sum\limits_i\left(ni-3\right)}$, so that $\sum\left(Z_i-\overline{Z}\right)\sqrt{ni-3}$; i=1,2,3,…,k are independent

standard normal variate .

Hence, $\sum_{i=1}^{k}(ni-3)\left(Z_i-\overline{Z}\right)^2$ is a Chi-square variate with (k-1) d.f. If χ^2 value thus obtained is greater than 5 percent value of χ^2 for (k-1) d.f, the null hypothesis of homogeneity of correlation coefficient is rejected. If not, the correlation coefficients are supposed to be homogenous in which case we combine the sample correlation coefficient to find the estimate P of the population correlation coefficient P
We have

$$\overline{Z}=\frac{1}{2}\log_e\left(\frac{1+P}{1-P}\right)\Rightarrow\left(1+P\right)=\left(1-P\right)e^{2\overline{Z}}$$

$$\Rightarrow 1+P=e^{2\overline{Z}}-Pe^{2\overline{Z}}$$

$$\Rightarrow P+Pe^{2\overline{Z}}=e^{2\overline{Z}}-1$$

$$\Rightarrow P\left(1+e^{2\overline{Z}}\right)=e^{2\overline{Z}}-1$$

$$\Rightarrow P=\frac{e^{2\overline{Z}}-1}{e^{2\overline{Z}}+1}=\tan^{-1}h\overline{Z}$$

$$R=\tan^{-1}h\overline{Z}$$

13.4 F – TEST FOR LINEARITY OF REGRESSSION:

F-Distribution:

If x and y are two independent chi-square variables with v_1 and v_2 degrees of freedom respectively. Then F-Statistic defined by

$$F=\frac{X/v_1}{Y/v_2}$$

A statistic F follows F-Distribution with (v_1, v_2) degrees of freedom will be denoted by $F\sim F(v_1,v_2)$.

Linear regression:

If the curve is a straight line, it is call the linear of regression and that is said to be linear regression between the variable otherwise regression said to be curvilinear.

Let us suppose that in the bivariate distribution (x_i,y_i); i=1,2,..,n, y is dependent variable and x is independent variable . let the line of regression of y on x be y = a+bx.

For a sample of size N arranged in n arrays, from a bivariate normal population, the test statistic foe testing the hypothesis of linearity of regression is

$$F=\frac{\eta^2-r^2}{1-\eta^2}.\frac{N-h}{h-2}$$

13.5 Variance stabilizing transformation

Variance-stabilizing transformations are mathematical functions applied to data to make the variance approximately constant when the original data exhibit heteroscedasticity, that is, when the variability changes with the mean. Many statistical methods, such as regression, ANOVA, and t-tests, assume that the variance of the observations is uniform across all levels of the predictor variables. When this assumption is violated, the results may become unreliable. To overcome this problem, variance-stabilizing transformations such as the logarithmic, square-root, reciprocal, and arcsine transformations are used depending on the nature of the data. For example, the square-root transformation is suitable for Poisson count data, the log transformation is used when variance increases proportionally to the square of the mean, and the arcsine transformation is applied to proportions. By stabilizing the variance and often reducing skewness, these transformations help the data better satisfy model assumptions and improve the accuracy and interpretability of statistical conclusions.

1. Square-root Transformation

Used when variance is proportional to the mean

$$Var(X) \alpha \ \mu$$

Typical for:

- Poisson counts (events, accidents, phone calls)

Transformation:

$$T(X) = \sqrt{X}$$

2. Log Transformation

Used when variance is proportional to the square of the mean

$$Var(X) \alpha \ \mu^2$$

Typical for:

- Right-skewed data
- Income
- Biological growth data

Transformation:

$$T(X) = \log(X)$$

3. Arcsine (Arcsin√p) Transformation

Used for proportions or percentages**:**

$$Var(p) = \frac{p(1-p)}{n}$$

Transformation:

$$T(p) = \arcsin\left(\sqrt{p}\right)$$

Used in:

- Binomial data
- Success–failure proportions

4. Reciprocal Transformation

Used when variance increases rapidly with mean:

$$T(X) = \frac{1}{X}$$

5. Box–Cox Transformation

A general method to determine the best transformation:

$$T(X) = \begin{cases} \frac{X^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ Log\ X, & \lambda = 0 \end{cases}$$

Advantages

- Makes variance constant → satisfies ANOVA/regression assumptions
- Reduces heteroscedasticity
- Improves normality of data
- Stabilizes residual patterns
- Helps remove skewness

Disadvantages

- Choice of transformation may not be obvious
- Interpretation becomes more difficult (in transformed units)
- Over- or under-transformation can distort results
- Back-transformation may introduce bias

13.6 TESTS OF SIGNIFICANCE FOR LARGE SAMPLES:

In section, we will discuss the test of significance when samples are large, we have seen that for large values of n, the number of trails, almost all the distributions. E.g.: Binomial, Poisson, Negative binomial etc…, are very closely approximated by normal distribution. Thus in this case we apply the normal i.e., which is based upon the following fundamental property (are property) of the normal probability curve.

$$\text{If } X \sim N(\mu,\sigma^2), \text{then } Z = \frac{x-\mu}{\sigma} = \frac{x-f(x)}{\sqrt{v(x)}} \sim N(0,1)$$

Thus from the normal probability, we have

$P(-3 \le Z \le 3) = 0.9973$ i.e., $P(|Z| < 3)$

$P(|Z| < 3) = 0.001$, $1 - P(|Z| < 3) = 0.0021$

i.e., in all probability we should expect a standard normal variate to lie between ± 3.

Also from the normal probability

tables, we get

$P(-1.96 \le Z \le 1.96) = 0.95$ i.e., $P(Z \le 1.96) = 0.95$

$P(Z > 1.96) = 1 - 0.95 = 0.05$

and $P(|Z| < 2.58) = 0.99$ $P(|Z| > 2.58) = 0.01$

thus the significant values of z at 5% and 1% levels of significance for a two trailed test are 1.96 and 2.58 respectively. Thus the steps to be used in the normal test is as follows

i) Compute the test statistic 'Z' under H_0

ii) If $|Z| > 3$, H_0 is always rejected.

iii) If $|Z| \le 3$, we test its significance at certain levels of significance, usually at 5% and sometimes at 1% level of significance. Thus for a two – tailed test if $|Z| > 1.96$, H_0 is rejected at 5% level of significance.

From the normal probability tables, we have

$P(Z > 1.645) = 0.5 - P(0 \le Z \le 1.645) = 0.5 - 0.45 = 0.05$

$P(Z > 2.33) = 0.5 - P(0 \le Z \le 2.33) = 0.5 - 0.49 = 0.01$

Hence for the single tailed test (Right tail (or) Left tail) we compare value of |Z| with 1.645 (at 5% level) and 2.33 (at 1% level) and accept otherwise reject H_0 accordingly.

13.6.1 COMPARISION OF χ^2 AND KOLMOGOROV SMIRNOV TEST:

We compare the two test of goodness of fit.

Chi – square test	Kolmogorov – Smirnov test
1) This groups the data and often loses information through grouping.	This treat individual sample observations directly.
2) This test is designed mostly for large samples.	Applicable even in the case of small samples.
3) It allows the estimation of unknown parameters in F_0. Then suggests a test with modified degrees of freedom.	These tests have no much provision.
4) This can be used in both the case when data are in the form of natural categorise (or) continuous.	This test essentially assumes the continuity of the parent CDF, therefore, it provides more refined analysis of data. It is conservative in the sense that it is more H_0 when applied in case F_x is not continuous.

13.7 CONCLUSION

The methods discussed in Sections **13.2 to 13.6** play a vital role in ensuring the reliability and accuracy of statistical inference. Bartlett's test safeguards analyses that rely on equal variances, while the Chi-square test for homogeneity of correlation coefficients helps compare relationships across different samples. The F-test for linearity ensures that regression models are correctly specified before interpretation or prediction. Variance-stabilizing transformations allow data to meet essential assumptions required by many statistical procedures. Tests of significance for large samples, grounded in the Central Limit Theorem, provide powerful tools for inference when sample sizes are sufficiently large.

Together, these techniques strengthen the foundations of applied statistics by validating assumptions, correcting irregularities in data, and enabling more precise decision-making. They collectively form an essential toolkit for researchers, analysts, and students working with real-world data.

13.8 SELF ASSESSMENT QUESTIONS

1. Explain the purpose of Bartlett's test and describe a situation where it is necessary to check the homogeneity of variances.
2. Using Fisher's Z-transformation, show how correlation coefficients from different samples can be compared for homogeneity.
3. Write the null and alternative hypotheses for testing the linearity of a regression model using the F-test.

4. What is a variance-stabilizing transformation? Give two practical examples where such transformations are applied.
5. Derive the Z-test statistic used to test the significance of a large sample mean.
6. Discuss the steps involved in testing the homogeneity of several independent correlation coefficients using the Chi-square test.

13.9 SUGGESTED READING BOOKS:

1. Statistical Inference by H.C, Saxena & Surendran
2. An outline of Statistical theory vol.2 by A.M. Goon and B. Das Gupta.
3. An Introduction to probability and Mathematical statistics by V.K. Rohatgi.
4. Mathematical Statistics- Parimal Mukopadhyay(1996), New central Book Agency (P)Ltd., Calcutra.

Dr. Syed Jilani

LESSON -14

SEQUENTIAL TESTS & SPRT

OBJECTIVES:

By the end of this lesson, students will be able to:

- ❖ Understand the fundamental concept of sequential hypothesis testing**,** and distinguish it from fixed-sample classical hypothesis testing.
- ❖ Describe the mechanism by which the SPRT makes real-time decisions by continuously evaluating accumulating data.
- ❖ Apply the correct notation for sequential tests, including likelihood ratios (L_n), decision boundaries $(A \text{ and } \mathrm{B})$, and error probabilities $(\alpha \text{ and } \beta)$.
- ❖ Explain and construct Sequential Probability Ratio Tests (SPRT) based on likelihood ratios for different statistical models.
- ❖ Work through examples of SPRT applications, such as in quality control or clinical trials, to determine when to accept the null hypothesis (H_0), reject the null hypothesis (accept (H_1), or continue sampling.

STRUCTURE:

14.1 INTRODUCTION:

In classical (fixed-sample) hypothesis testing, the sample size *n* is predetermined before data collection. After all observations are collected, the test statistic is computed and a decision is made.

However, this fixed-sample approach may be inefficient in many real-world situations where:

- Data arrives sequentially over time (quality control, clinical trials, industrial processes).
- Observations are costly**,** and fewer samples should be used when evidence is strong.
- Early decisions are required to reduce time, cost, or risk**.**

To address these issues, Sequential Analysis was developed - primarily by Abraham Wald (1940s) **-** as a methodology where data is evaluated as it is collected, and decisions can be made at any stage**.**

14.2 NOTATION OF SEQUENTIAL TESTS:

A Sequential Test is a method of hypothesis testing in which:

- Observations are taken one at a time (or in small groups).
- After each observation, a decision is made whether to:
- Accept the null hypothesis (H_0).
- Reject the null hypothesis (Accept the alternative hypothesis, H_1).
- Continue sampling (take another observation).

Definition

A sequential test is a statistical procedure where the sample size is not fixed in advance**.** Instead, the number of observations required is a random variable determined by a stopping rule.

Key Features

- Sample size is variable and data-dependent.
- Decisions are dynamic and updated step-by-step.
- The test typically requires fewer observations than fixed-sample tests.
- Optimality results exist (Ex: SPRT minimizes the expected sample size).

Applications:

Sequential testing is widely used across various fields:

- **Clinical Trials:** Patients are enrolled sequentially, and interim analyses are conducted to determine if a new drug is effective, ineffective (futility), or if the trial should continue. This can stop an ineffective trial early, saving costs and ethical concerns.
- **Quality Control/Manufacturing:** In a production line, sequential sampling can monitor the defect rate. Testing items one by one allows for immediate intervention if a defect threshold is crossed, minimizing waste.
- **Ecological Monitoring:** Assessing pest counts or environmental conditions (like radiation leaks) sequentially enables rapid response to hazardous levels.
- **Finance:** Sequential analysis can be applied to market trends, allowing for quick investment decisions (buy/sell) based on fluctuating market data to maximize returns or minimize losses.

14.3 SEQUENTIAL PROBABILITY RATIO TEST (SPRT):

So far, you studied the hypothesis testing after observing the entire data. Alternatively, in various practical experiments, the data is analysed sequentially as they are collected, and further sampling is stopped once there is already enough evidence for making a conclusion. Sequential testing may save experimental costs ad time. Furthermore, in some situations, the decision must be made in "real-time." For example, a quality control engineer controls a certain chemical technological process that requires keeping the temperature at a fixed level. The temperature is measured every 5 minutes, and the engineer must decide whether the process is still operating properly or the temperature has been changed and the process should be terminated. This is also an example, where sequential hypothesis testing is essential.

In traditional hypothesis testing, we fix the sample size in advance. As you have seen in the Neyman-Pearson lemma for testing a simple null hypothesis against a simple alternative hypothesis. Abraham Wald in 1945 (during World War II, Wald was a member of

the Statistical Research Group at Columbia University, where he applied his statistical skills to various wartime problems. They included methods of sequential analysis and sampling inspection) proposed an extension of the Neyman-Pearson lemma in which both Type-I error and Type-II error are fixed and the sample size is not fixed and considered as a random variable. It is a statistical method for hypothesis testing in which data is evaluated sequentially as it is collected, instead of requiring a fixed sample size. This means that rather than waiting until a pre-determined number of observations have been collected, decisions can be made at any point in the sampling process based on the accumulated evidence. This technique is called Sequential Probability Ratio Test (SPRT). This approach is more efficient than traditional hypothesis testing methods because it minimises the average sample size (ASN) needed to reach a conclusion while maintaining the desired levels of accuracy. Due to this, sequential probability ratio test has found applications in a wide range of fields, including medical, quality control, industrial inspections, finance, defence systems, etc.
Let us discuss the formal statement of the sequential probability ratio test.

Let $X_1, X_2, ...$ be a sequence of independent observations taken from a population whose probability density/ mass function is $f(X;\theta)$ which depends on a parameter θ which takes one of the two values θ_0 or θ_1. Suppose we want to test a simple null hypothesis

$H_0 : \theta = \theta_0$

Against a simple alternative hypothesis

$H_1 : \theta = \theta_1$

and $L(X;\theta_0)$ *and* $L(X;\theta_1)$ are the likelihood functions of the sample observations under the null hypothesis $H_0 : \theta = \theta_0$ and alternative hypothesis $H_1 : \theta = \theta_1$ respectively, then at each stage m, the likelihood ratio is calculated as

$$\lambda_m = \frac{L_m(X;\theta_1)}{L_m(X;\theta_0)} \text{ for every m = 1,2,.....}$$

$$\lambda_m = \frac{f(X_1;\theta_1) f(X_2;\theta_1) ..... f(X_m;\theta_1)}{f(X_1;\theta_0) f(X_2;\theta_0) ..... f(X_m;\theta_0)}$$

$$\lambda_m = \prod_{i=1}^{m} \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}$$

The decision regarding the null hypothesis (H_0) is based on comparing λ_m with two pre-determined constants A *and* B $(B < A)$.

Thus, the sequential probability ratio test for testing a simple null hypothesis H_0 against a simple alternative hypothesis H_1 is defined as follows:

(i) If $\lambda_m \geq A$, then we stop the process and reject the null hypothesis H_0.

(ii) If $\lambda_m \leq B$, then we stop the process and accept the null hypothesis H_0.

(iii) $B < \lambda_m < A$, then we continue sampling by taking an additional observation.

We determine the constants A and B so that the sequential probability ratio test will have pre assigned α *and* β. The actual determination of both A and B is in general quite difficult. Therefore, we use approximate values of A and B which are given as follows:

$$A = \frac{1-\beta}{\alpha} \text{ and } B = \frac{\beta}{1-\alpha}$$

For the computational point of view, we deal with $\log(\lambda_m)$ instead of λ_m, therefore, we calculate

$$\log(\lambda_m) = \log \prod_{i=1}^{m} \frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}$$

$$\log(\lambda_m) = \sum_{i=1}^{m} \log\left\{\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}\right\}$$

$$\log(\lambda_m) = \sum_{i=1}^{m} Z_i$$

$$where\ Z_i = \log\left\{\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}\right\}$$

We can define the sequential probability ratio test in terms of Z as follows:

(i) If $\sum_{i=1}^{m} Z_i \geq \log(\mathrm{A})$, then we stop the process and reject the null hypothesis H_0.

(ii) If $\sum_{i=1}^{m} Z_i \leq \log(\mathrm{B})$, then we stop the process and accept the null hypothesis H_0.

(iii) If $\log(\mathrm{B}) < \sum_{i=1}^{m} Z_i < \log(\mathrm{A})$, then we continue sampling by taking an additional observation.

Since this test is given by Wald so it is also known as a sequential Wald test. Note that unlike the "standard" hypotheses testing setup, where there are two possible decisions: to accept or to reject the null, in sequential testing at any given time there is an additional "neutral" option: not to decide yet and wait for the next observation. This process continues until we reach one of the two thresholds, ensuring an efficient decision with the minimum expected sample size compared to fixed-sample tests. It can be shown that the stopping time T is finite with probability one, that is, $P(T < \infty) = 1$.

Note 1: The proof of this theorem is beyond the scope of this course.

Procedure to Apply Sequential Probability Ratio Test:

The sequential probability ratio test is an extension of the Neyman-Pearson lemma so it has almost the same procedure as the Neyman-Pearson lemma but for simplicity, we use it in terms of Z. Let us discuss as follows:

Step 1: First of all, we identify a simple null hypothesis (H_0) and a simple alternative hypothesis (H_1) as we formulated in the Neyman-Pearson lemma, that is, both hypotheses specify exact parameter values.

$H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$,

Step 2: After identifying the simple null and alternative hypotheses, we find the probability density/mass function of the data. Then we compute the probability density/mass function under the null and alternative hypotheses as $f(X_i;\theta_0)$ and $f(X_i;\theta_1)$.

Step 3: After that, we compute Z_i as follows:

$$Z_i = \log\left\{\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}\right\}$$

and then we find the sum

$$\sum_{i=1}^{m} Z_i = \sum_{i=1}^{m} \log\left\{\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}\right\}$$

Step 4: Finally, we define the sequential probability ratio test in terms of Z as follows:

(i) If $\sum_{i=1}^{m} Z_i \geq \log(\mathrm{A})$, then we stop the process and reject the null hypothesis H_0.

(ii) If $\sum_{i=1}^{m} Z_i \leq \log(\mathrm{B})$, then we stop the process and accept the null hypothesis H_0.

(iii) If $\log(\mathrm{B}) < \sum_{i=1}^{m} Z_i < \log(\mathrm{A})$, then we continue sampling by taking an additional observation.

After understanding the procedure to apply the sequential probability ratio test, let us discuss how to apply it with the help of an example.

14.4 Example:

An environmental monitoring agency of a city regularly measures the Air Quality Index (AQI) to assess pollution levels. On the basis of past data, it is observed that the AQI follows a normal distribution with a mean AQI $\mu = 50$ and a known standard deviation $\sigma = 10$. The agency observed that due to increasing vehicles and construction sites, the AQI has increased.

(i) Derive sequential probability ratio test for testing the hypothesis $H_0 : \mu = \mu_0$=50 against $H_1 : \mu = \mu_1 = 65$ for given $\alpha = 0.05$ and $\beta = 0.1$.

(ii) If the agency takes observations/samples 52, 55, 58, 63, 66, 70, 74 sequentially then show the step-by-step decision using SPRT.

Solution: The AQI follows a normal distribution, and we know that the probability density function of the normal distribution with mean μ and variance σ^2 as

$$f(X;\mu,\sigma^2) = \frac{1}{\sqrt{2\prod\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Since the AQI follows a normal distribution with a known standard deviation, therefore, we can write the probability density function of the normal distribution as follows:

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

We now compute the pdf under the null hypothesis (H_0) and alternative hypothesis (H_1) as follows:

$$f(X;\mu_0) = \frac{1}{\sqrt{2\prod\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu_0)^2}$$

Similarly,

$$f(X;\mu_1)=\frac{1}{\sqrt{2\prod\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu_1)^2}$$

Therefore, we can compute $Z_i=\log\left\{\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}\right\}$ as follows:

$$Z_i=\log\left\{\frac{f(X_i;\mu_1)}{f(X_i;\mu_0)}\right\}=\log\left\{\frac{\frac{1}{\sqrt{2\prod\sigma^2}}e^{-\frac{1}{2\sigma^2}(x_i-\mu_1)^2}}{\frac{1}{\sqrt{2\prod\sigma^2}}e^{-\frac{1}{2\sigma^2}(x_i-\mu_0)^2}}\right\}$$

We now try to simplify it as follows:

$$Z_i=-\frac{1}{2\sigma^2}(X_i-\mu_1)^2+\frac{1}{2\sigma^2}(X_i-\mu_0)^2$$

$$Z_i=-\frac{1}{2\sigma^2}\left\{X_i^2+\mu_1^2-2X_i\mu_1-X_i^2-\mu_0^2+2X_i\mu_0\right\}$$

$$Z_i=-\frac{1}{2\sigma^2}\left\{\mu_1^2-\mu_0^2-2X_i\mu_1+2X_i\mu_0\right\}$$

$$Z_i=-\frac{1}{2\sigma^2}\left\{(\mu_1+\mu_0)(\mu_1-\mu_0)-2X_i(\mu_1-\mu_0)\right\}$$

$$Z_i=\frac{(\mu_1-\mu_0)}{\sigma^2}\left\{X_i-\frac{(\mu_1+\mu_0)}{2}\right\}$$

Hence,

$$\sum_{i=1}^{m}Z_i=\frac{(\mu_1-\mu_0)}{\sigma^2}\left\{\sum_{i=1}^{m}X_i-\frac{m(\mu_1+\mu_0)}{2}\right\}$$

Hence, the SPRT for testing $H_0:\mu=\mu_0$=50 against $H_1:\mu=\mu_1=65$ is given as

(i) Reject H_0 if

$$\sum_{i=1}^{m}Z_i\geq\log\left(\frac{1-\beta}{\alpha}\right)$$

$$\frac{(\mu_1-\mu_0)}{\sigma^2}\left\{\sum_{i=1}^{m}X_i-\frac{m(\mu_1+\mu_0)}{2}\right\}\geq\log\left(\frac{1-\beta}{\alpha}\right)$$

We now try to simplify it, multiplying the inequality by $\frac{\sigma^2}{(\mu_1-\mu_0)}>0$

$(\because\mu_1>\mu_0)$, *we get*:

$$\sum_{i=1}^{m}X_i-\frac{m(\mu_1+\mu_0)}{2}\geq\frac{\sigma^2}{(\mu_1-\mu_0)}\log\left(\frac{1-\beta}{\alpha}\right)$$

Now, we try to simplify in terms of the $\sum_{i=1}^{m}X_i$ as follows:

$$\sum_{i=1}^{m}X_i\geq\frac{\sigma^2}{(\mu_1-\mu_0)}\log\left(\frac{1-\beta}{\alpha}\right)+\frac{m(\mu_1+\mu_0)}{2}$$

(ii) Accept H_0 if

$$\sum_{i=1}^{m} Z_i \le \log\left(\frac{\beta}{1-\alpha}\right)$$

$$\frac{(\mu_1-\mu_0)}{\sigma^2}\left\{\sum_{i=1}^{n} X_i - \frac{m(\mu_1+\mu_0)}{2}\right\} \le \log\left(\frac{\beta}{1-\alpha}\right)$$

$$\sum_{i=1}^{m} X_i \le \frac{\sigma^2}{(\mu_1-\mu_0)}\log\left(\frac{\beta}{1-\alpha}\right) + \frac{m(\mu_1+\mu_0)}{2}$$

(iii) Continue taking additional observations as long as

$$\log\left(\frac{\beta}{1-\alpha}\right) < \sum_{i=1}^{m} Z_i < \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\log\left(\frac{\beta}{1-\alpha}\right) \le \frac{(\mu_1-\mu_0)}{\sigma^2}\left\{\sum_{i=1}^{m} X_i - \frac{m(\mu_1+\mu_0)}{2}\right\} \le \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\frac{\sigma^2}{(\mu_1-\mu_0)}\log\left(\frac{\beta}{1-\alpha}\right) + \frac{m(\mu_1+\mu_0)}{2} < \sum_{i=1}^{m} X_i < \frac{\sigma^2}{(\mu_1-\mu_0)}\log\left(\frac{1-\beta}{\alpha}\right) + \frac{m(\mu_1+\mu_0)}{2}$$

Here, it is given that

$\mu_0 = 50,\ \mu_1 = 65,\ \sigma$=10, α=0.05, β=0.1, therefore we calculate

$$A = \frac{1-\beta}{\alpha} = \frac{1-0.1}{0.05} = 18$$

$$\log(A) = \log(18) = 2.89$$

Similarly,

$$B = \frac{\beta}{1-\alpha} = \frac{0.1}{1-0.05} = 0.105$$

$$\log(B) = \log(0.105) = -2.25$$

Therefore, we the SPRT is give as follows:

(i) Reject H_0 if

$$\sum_{i=1}^{m} X_i \ge \frac{\sigma^2}{(\mu_1-\mu_0)}\log\left(\frac{1-\beta}{\alpha}\right) + \frac{m(\mu_1+\mu_0)}{2}$$

$$\sum_{i=1}^{m} X_i \ge \frac{100}{(65-50)}(2.89) + \frac{m(65+50)}{2} = 19.27 + 57.5m$$

(ii) Accept H_0 if

$$\sum_{i=1}^{m} X_i \le \frac{\sigma^2}{(\mu_1-\mu_0)}\log\left(\frac{\beta}{1-\alpha}\right) + \frac{m(\mu_1+\mu_0)}{2}$$

$$\sum_{i=1}^{m} X_i \le \frac{100}{(65-50)}(-2.25) + \frac{m(65+50)}{2} = -15 + 57.5m$$

(iii) Continue taking additional observations as long as

$$-15 + 57.5m < \sum_{i=1}^{m} X_i < 19.27 + 57.5m$$

We now show the step-by-step decision using SPRT.

$$Z_i = \frac{(\mu_1 - \mu_0)}{\sigma^2}\left\{X_i - \frac{(\mu_1 + \mu_0)}{2}\right\} = \frac{(65-50)}{100}\left\{X_i - \frac{115}{2}\right\} = 0.15X_i - 8.625$$

We calculate $\sum_{i=1}^{m} Z_i$ at each step and compare it with log(A) and log(B) and take the decision about the null hypothesis H0 as shown in the following table:

Sample	X	$Z_i = 0.15X_i - 8.625$	Cumulative $\sum_{i=1}^{m} Z_i$	Decision
1	52	-0.825	-0.825	Continue
2	55	-0.375	-1.20	Continue
3	58	0.075	-1.125	Continue
4	63	0.825	-0.30	Continue
5	66	1.275	0.975	Continue
6	70	1.875	2.85	Continue
7	74	2.625	5.475	Reject H_0

Hence, the SPRT method efficiently detected an increase in AQI with only 7 observations/samples instead of using a fixed-sample test (which may require 30 or more than 30 observations).

14.5 PROPERTIES OF SEQUENTIAL PROBABILITY RATIO TEST:

The sequential probability ratio test is a statistical method for hypothesis testing in which data is evaluated sequentially as it is collected, instead of requiring a fixed sample size. It has several key properties that make it widely applicable in hypothesis testing. Some important properties of it are as follows:

1. The sequential probability ratio test evaluates data as it is collected and the test stops as soon as sufficient evidence is found to accept or reject the hypothesis.
2. The SPRT minimizes the expected number of observations while maintaining pre-specified Type-I and Type-II errors.
3. It is the most efficient sequential test under certain conditions.
4. The test is designed to control the probabilities of Type-I and Type-II errors. The thresholds A and B are chosen based on desired errors.
5. The test often reaches a decision faster than fixed-sample tests, reducing costs and time.

14.6 SUMMARY:

Sequential Probability Ratio Testing is a powerful and efficient method for hypothesis testing when observations arrive one at a time. Using likelihood ratios and decision boundaries, SPRT achieves rapid and accurate decisions while controlling Type I and Type II errors. Its optimality in terms of the expected sample size makes it superior to traditional fixed-sample procedures. The theory-supported by OC and ASN functions-demonstrates that sequential tests can save time, cost, and resources in many real-world applications. Overall, SPRT provides a scientific, data-driven, and resource-efficient framework for continuous decision-making in statistics.

14.7 KEY WORDS:

The most important keywords for this topic are:

- Sequential Test (or Sequential Analysis)
- Sequential Probability Ratio Test (SPRT)
- Hypothesis Testing (H_0 vs. H_1)
- Likelihood Ratio (L_n)
- Acceptance/Rejection Boundaries (A *and* B)
- Continuation Region

14.8 SELF-ASSESSMENT QUESTIONS:

1. What is meant by sequential analysis*?* How does it differ from classical (fixed-sample) hypothesis testing?
2. Explain why sequential tests can be more efficient than fixed-sample tests.
3. Give two real-life situations where sequential testing is preferable to fixed-sample tests.
4. Define a stopping rule (T) *and* a decision rule (d*)* in a sequential test.
5. What do the quantities α and β represent? How are they used in designing sequential tests?
6. Write the general structure of a sequential test procedure using likelihood ratios.
7. State the decision rule of SPRT using likelihood ratio boundaries A *and* B .
8. Why is SPRT considered the most efficient sequential test?
9. Describe the continuation region in SPRT.
10. Construct an SPRT for testing $H_0 : p = p_0$ vs. $H_1 : p = p_1$ for Bernoulli trials.

14.9 SUGGESTED READINGS:

- Mathematical Statistics- Parimal Mukopadhyay(1996), New Central Book Agency (P)Ltd., Calcutta
- Statistical Inference by H.C, Saxena & Surendran
- An outline of Statistical Theory, vol.2 by A.M. Goon and B. Das Gupta
- An Introduction to probability and Mathematical Statistics by V.K. Rohatgi
- Wald, A. – Sequential Analysis
- Ghosh, J.K. – Sequential Analysis
- Hogg, McKean & Craig – Introduction to Mathematical Statistics
- Casella & Berger – Statistical Inference
- Rao, C.R. – Linear Statistical Inference and Its Applications.

Dr. K. Kalyani

LESSON -15

WALD'S FUNDAMENTAL IDENTITY & RELATIONSHIP BETWEEN A, B, α and β

OBJECTIVES:

By the end of this lesson, students will be able to:

- Understand Wald's Fundamental Identity in the context of sequential analysis.
- Explain the meaning of Type I error (α), Type II error (β), and decision boundaries A and B.
- Establish the relationship between Wald's decision boundaries and the error probabilities.
- Apply Wald's identity to derive approximations for A and B in the Sequential Probability Ratio Test (SPRT).
- Interpret how error probabilities influence the design of sequential tests.
- Use Wald's identity to evaluate stopping rules in practical testing situations.

STRUCTURE:

15.1 INTRODUCTION:

Sequential analysis is a powerful methodology in statistical inference where data are evaluated as they are collected, and decisions are made at any stage of sampling. Unlike fixed-sample hypothesis tests, sequential tests allow early acceptance or rejection of a hypothesis, often reducing the average sample size while maintaining desired error probabilities. Among the foundational contributions to sequential testing is the work of Abraham Wald, who developed the Sequential Probability Ratio Test (SPRT)-a test proven to be optimal in terms of expected sample size.

A key component of Wald's theory is the likelihood ratio, which is monitored continuously during sequential testing. Decisions are made by comparing this likelihood ratio to two constants, A and B, which define the upper and lower stopping boundaries. The

selection of these boundaries is crucial because it directly controls the Type I error (α) and Type II error (β) of the test.

Wald introduced a fundamental identity that links the expected values of the likelihood ratio under different hypotheses to the error probabilities. This leads to the well-known approximate relationships:

$$A \approx \frac{1-\beta}{\alpha},\ \mathrm{B} \approx \frac{\beta}{1-\alpha}$$

These relationships form the practical basis for designing an SPRT. By choosing α and β (the desired error probabilities), one can compute A and B, thereby establishing decision rules for the sequential test.

This lesson explains Wald's fundamental identity and shows how it provides the connection between decision boundaries (A, B) and error probabilities (α, β), forming the core of sequential test design.

15.1.1 OVERVIEW OF SEQUENTIAL HYPOTHESIS TESTING:

Sequential hypothesis testing is designed to test:

$$H_0 \text{ vs. } H_1$$

by observing data one point at a time. After each new observation, a likelihood ratio is computed:

$$\Lambda_n = \frac{L_1\left(X_1,.....X_n\right)}{L_0\left(X_1,.....X_n\right)},$$

and compared with two boundaries:

- Upper boundary (A): Reject H_0
- Lower boundary (B): Accept H_0
- Intermediate region: Continue sampling

This dynamic decision-making leads to tests that often require many fewer observations than fixed-sample tests with the same error probabilities.

Sequential tests are optimal in many settings and provide strong control over:

- Type I error (α): Probability of rejecting H_0 when it is true
- Type II error (β): Probability of accepting H_0 when H_1 is true

These error probabilities determine the stopping boundaries A and B.

15.1.2 ROLE OF WALD IN SEQUENTIAL ANALYSIS:

Abraham Wald is the founder of modern sequential analysis. His contributions include:

1. Development of the Sequential Probability Ratio Test (SPRT)

- Wald introduced SPRT as a test that compares the likelihood ratio to fixed thresholds A and B.
- He proved that SPRT is optimal**,** meaning it minimizes the expected sample size among all tests satisfying the same error constraints.

2. Wald's Fundamental Identity

- Wald derived a key identity relating the expected value of the likelihood ratio to the error probabilities.
- This identity provides the basis for determining the decision boundaries A and B.

3. Linking A, B to α and β

Using Wald's reasoning:

$$A \approx \frac{1-\beta}{\alpha},\ \mathrm{B} \approx \frac{\beta}{1-\alpha}.$$

This remarkable relationship allows practitioners to **choose α and β first**, then derive A and B to construct the test.

4. Establishing the theory of optimal sequential tests

- Wald's work laid the theoretical foundation for sequential designs used today.
- His methods influence sequential clinical trials, quality control charts, and adaptive algorithms.

15.2 WALD'S FUNDAMENTAL IDENTITY:

Consider the SPRT for testing H_0 that the probability distribution of X is given by $f(X,\theta_0)$ against the alternative hypothesis H_1 that the probability distribution of X is given by $f(X,\theta_1)$.

$$Let\ Z = \log\frac{f(X,\theta_1)}{f(X,\theta_0)}$$

$$Z_i = \log\frac{f(X_i,\theta_1)}{f(X_i,\theta_0)}$$

Where X_i denote the ith observation on X then the SPRT Procedure is given below.

i) Continue observation taking as $\log B < Z_1 + Z_2 + ... + Z_m < \log A$. Where A and B (B < A) are constants determined before the experiment start.

ii) Accept H_0 when $\sum_{i=1}^{m} Z_i \le \log(\mathrm{B})$

iii) Reject H_0 when $\sum_{i=1}^{m} Z_i \geq \log(\mathrm{A})$

Let us denote 'n' the number of observations required by the tests clearly 'n' is a random variable.

Let 'D' be the subset of the random variable such that $\mathrm{E}\left(e^{Zt}\right) = M(t)$ exists, and these are finite for any point t and D.

Consider the following identity

$$\mathrm{E}\left[e^{S_n t+(S_N-S_n)t}\right] = E\left[e^{S_N t}\right] = \left[M(t)\right]^M \text{.....(1)}$$

Where

$$S_n = Z_1 + Z_2 + \ldots\ldots + Z_n$$
$$S_N = Z_1 + Z_2 + \ldots\ldots + Z_N$$

and N is any positive integer.

Let P_n denote the probability that $n \leq N$ for any random variable μ. Let $(E_N)\mu$ denote S the conditional expected value of 'μ' and restriction $n \leq N$ and Let $E_N^*(\mu)$ denote the conditional expected value of 'μ' and under the restriction when $n > N$.

$$P_N E_N\left[e^{S_n t+(S_N-S_n)t}\right] + (1-P_N)E_N^*\left(e^{S_N t}\right)E\left[M(t)\right]^M \text{.....(2)}$$

The expression $S_N - S_n$ is independent of S_n

We have

$$E_N\left[e^{S_n t+(S_N-S_n)t}\right] = E_N\left[e^{S_n t}\left[M(t)\right]^{N-n}\right] \text{.....(3)}$$

$$P_N E_N\left[e^{S_n t}\left[M(t)\right]^{N-n}\right] + (1-P_N)E_N^*\left(e^{S_N t}\right) = \left[M(t)\right]^N \text{.....(4)}$$

dividing $\left[M(t)\right]^N$ on both sides we get

$$\frac{P_N E_N\left[e^{S_n t}\left[M(t)\right]^{N-n}\right]}{\left[M(t)\right]^N} + \frac{(1-P_N)E_N^*\left(e^{S_N t}\right)}{\left[M(t)\right]^N} = \frac{\left[M(t)\right]^N}{\left[M(t)\right]^N}$$

$$P_N E_N\left[e^{S_n t}\left[M(t)\right]^{-n}\right] + \frac{(1-P_N)E_N^*\left(e^{S_N t}\right)}{\left[M(t)\right]^N} = 1 \text{.....(5)}$$

Let 'D' be the set of complex plain which $\left[M(t)\right] > 1$. Let 'D' denote the common part of subsets D^1 and D^{11} SPRT eventually terminates with probability one.

$$P_N = 1$$

$$1 - P_N = 1$$

$$\lim_{N\to\infty}(1-P_N)\frac{E_N^*\left(e^{S_N t}\right)}{\left[M(t)\right]^N} = 0 \text{.....(6)}$$

$$\therefore \lim_{N\to\infty} P_N . E_N\left[e^{S_n t}\left[M(t)\right]^{-n}\right] = E\left[e^{S_n t}\left[M(t)\right]^{-n}\right] \text{.....(7)}$$

Substitute equation (6) & (7) in equation (5)

we obtain the fundamental identity

$$E\left[e^{S_n t}\left[M(t)\right]^{-n}\right].$$

For any point 't' in the set D that is wald's fundamental identity.

15.3 ERROR PROBABILITIES AND DECISION BOUNDARIES:

The SPRT operates by calculating the likelihood ratio (or its logarithm) as data accumulates and comparing it to two pre-defined boundaries,

Likelihood Ratio (L_n): At step n, $L_n = \frac{P_1(X_1,.....X_n)}{P_0(X_1,.....X_n)}$.

❖ **Decision Rule:**

If $L_n \geq A$, stop and accept H_1 $(reject\ H_0)$.

If $L_n \leq B$, stop and accept H_0 $(reject\ H_1)$.

If $B < L_n < A$, continue sampling.

The boundaries A and B are determined by the desired Type I error rate (α, the probability of rejecting H_0 when it is true) and Type II error (β, the probability of accepting H_0 when H_1 is true).

15.4 Example 1 :

Consider testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, for a normal distribution with known variance. The log-likelihood ratio at stage n often simplifies to a cumulative sum related to the observations.

The boundaries can be set using the approximations $A \approx \frac{1-\beta}{\alpha}$ and $B \approx \frac{\beta}{1-\alpha}$. The test then proceeds by plotting the cumulative log-likelihood ratio against the sample number n and stopping when the path crosses either the upper boundary $l_n(A)$ or the lower boundary $l_n(B)$.

Example 2 :

Consider testing the mean of a normal distribution with known variance σ^2 :

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu = \mu_1,$$

For a sample $X_1, X_2, ..., $ the log-likelihood ratio per observation is:

$$\log \frac{f(X_i / \mu_1)}{f(X_i / \mu_0)} = \frac{\mu_1 - \mu_0}{\sigma^2}\left(X_i - \frac{\mu_1 + \mu_0}{2}\right).$$

After n observations:

$$\log \Lambda_n = \sum_{i=1}^{n} \frac{\mu_1 - \mu_0}{\sigma^2}\left(X_i - \frac{\mu_1 + \mu_0}{2}\right).$$

Stop as soon as

$$\log B \le \log \Lambda_n \le \log A$$

is violated.

This example illustrates how sequential testing accumulates evidence until the likelihood ratio crosses a boundary.

15.5 RELATIONSHIP BETWEEN A, B, α, β :

A sample $(X_1, X_2, ..., X_m)$ leads to acceptance of $H_0 : \theta = \theta_0$ if

$$B < \frac{L_1 m}{L_0 m} = \frac{f(X_1, \theta_1), ...f(X_m, \theta_1)}{f(X_1, \theta_0), ...f(X_m, \theta_0)} < A.....(1)$$

$$for\ m = 1, 2, ..., n-1$$

$$and\ \frac{L_1 n}{L_0 n} \le B.....(2)$$

and the sample leads to rejection of H_0.

$$If\ B < \frac{L_1 m}{L_0 m} < A, for\ m = 1, 2, ..., n-1$$

$$and\ \frac{L_1 n}{L_0 n} \ge A.....(3)$$

From (3) we have $L_1 n \ge A\ L_0 n$

$$1 - \beta \ge A\ \alpha.....(4)$$

$$A \le \frac{1-\beta}{\alpha}.....(5)$$

Thus $\frac{1-\beta}{\alpha}$ is an upper limit of A

Similarly from equation (2) we have $L_1 n \le B\ L_0 n$

$$\beta \le B(1-\alpha)$$

$$B \ge \frac{\beta}{1-\alpha} \text{.....(6)}$$

Example: Find the SPRT Procedure with stopping bounds $B < 1 < A$ for testing $H_0 : P = P_0$ against $H_1 : P = P_1$ when X has the p.d.f. given by

$$f(X,P) = \begin{cases} P^x.(1-P)^{1-X} & \text{if X=0,1} \\ 0 & \text{otherwise} \end{cases}$$

Solution: The densities function under H_0 and H_1 is

$$Z_i = \log \frac{f(X_i, P_1)}{f(X_i, P_0)} = \begin{cases} \log \frac{P_1}{P_0} & \text{if X=1} \\ \log \frac{1-P_1}{1-P_0} & \text{if X=0} \end{cases}$$

If 'r' is the no.of sequence of first n-observations, then

$$\sum_{i=1}^{n} Z_i = r \log \frac{P_1}{P_0} + (n-r) \log \frac{1-P_1}{1-P_0}$$

i) *If* $\sum_i Z_i \ge \log A$, reject H_0

ii) *If* $\sum_i Z_i \le \log B$, accept H_0

iii) If $\log B < \sum_i Z_i < \log A$

continuous sampling by testing one more observation.

15.6 APPLICATIONS IN SPRT:

The SPRT's ability to reduce the average sample size makes it valuable in various fields, especially where data collection is costly or time-consuming.
Key applications include:

- **Quality Control/Industrial Inspection:** Efficiently checking if a batch of products meets standards.
- **Clinical Trials/Medical Research:** Stopping a trial early if a new treatment is clearly effective (or ineffective), saving time and resources.
- **A/B Testing (Online Experiments):** Quickly identifying whether a website change improves a conversion rate.
- **Reliability Testing:** Lifetime testing of components and Accelerated failure tests
- **Finance and Econometrics**.

15.7 SUMMARY:

The Sequential Probability Ratio Test (SPRT), developed by Abraham Wald, is a powerful methodology for hypothesis testing where observations are evaluated sequentially. By using the likelihood ratio and Wald's decision boundaries, SPRT provides an optimal balance between accuracy and efficiency, minimizing the expected sample size while controlling Type I and Type II error probabilities. Wald's fundamental identity serves as the theoretical foundation of the test. The flexibility and efficiency of SPRT make it useful in quality control, medical testing, real-time signal detection, and other fields where quick decisions are crucial.

15.8 KEY WORDS:

- Sequential Analysis
- SPRT
- Likelihood Ratio
- Wald's Fundamental Identity
- Decision Boundaries:
  Alpha Error (α), Beta Error (β)
- Sequential Testing

15.9 SELF-ASSESSMENT QUESTIONS:

1. Explain Wald's contribution to sequential analysis. Write Wald's fundamental identity.
2. What are the decision boundaries A and B in SPRT? Describe the role of error probabilities α and β in SPRT.
3. State Wald's Fundamental Identity and explain the conditions under which it holds.
4. Derive the decision rules for SPRT using the likelihood ratio.
5. Explain the relationship between A, B, α, and β.
6. Describe the OC and ASN functions in the context of SPRT.
7. Discuss real-world applications of SPRT with examples.
8. Solve a full example of SPRT for a Bernoulli or normal distribution case.

15.10 SUGGESTED READINGS:

1 An outline of Statistical Theory, vol.2 by A.M. Goon and B. Das Gupta
2 An Introduction to probability and Mathematical Statistics by V.K. Rohatgi
3 Wald, A. – Sequential Analysis
4 Ghosh, J.K. – Sequential Analysis
5 Hogg, McKean & Craig – Introduction to Mathematical Statistics
6 Mathematical Statistics- Parimal Mukopadhyay(1996), New Central Book Agency (P)Ltd., Calcutta.

Dr. K. Kalyani

LESSON -16

OPERATING CHARACTERISTIC (OC) AND AVERAGE SAMPLE NUMBER (ASN) FUNCTIONS IN SPRT

OBJECTIVES:

By the end of this lesson, students will be able to:

- Explain the concepts of Operating Characteristic (OC) and Average Sample Number (ASN) in the context of SPRT.
- Understand how OC and ASN functions evaluate test performance in sequential hypothesis testing.
- Derive the OC function using Wald's fundamental identity.
- Obtain the ASN function for SPRT under both hypotheses (H_0) and (H_1)
- Interpret decision boundaries, Type I and II error probabilities in terms of OC and ASN.
- Apply OC and ASN properties to design efficient SPRTs in practical scenarios.

STRUCTURE:

16.1 INTRODUCTION:

The Sequential Probability Ratio Test (SPRT), introduced by Abraham Wald, is one of the most powerful and efficient procedures for sequential hypothesis testing. Unlike fixed-

sample tests, SPRT evaluates data as it is collected and terminates the experiment as soon as sufficient evidence is accumulated.

Two major functions used to study and evaluate the performance of an SPRT are:

- **Operating Characteristic (OC) function**
- **Average Sample Number (ASN) function**

The **OC function** gives the probability of accepting the null hypothesis for different parameter values, while the **ASN function** indicates the **expected number of observations** required before the test stops.

These functions help researchers understand the efficiency, error control, and behaviour of the SPRT under various conditions.

16.2 OPERATING CHARACTERISTIC (OC) FUNCTION:

In the sequential probability ratio test, decisions are made sequentially based on sequential data. To analyse the performance of the sequential probability ratio test, we use the operating characteristic (OC) function, which measures the probability of accepting the null hypothesis (H_0) for different values of the parameter being tested. The OC function helps us to know how well the test distinguishes between null and alternative hypotheses and gives the probability of making the correct or incorrect decision at different parameter values.

We denote the operating characteristic function as $P(\theta)$ and define it as follows:

$P(\theta)$ = Probability of accepting H_0 when θ is the true value of the parameter

$$P(\theta) = P(Accepting\ H_0 / \theta)$$

Also, by the definition of the power function of a test, we have

$\beta(\theta)$ = Probability of rejecting H_0 when θ is the true value

Therefore,

$$P(\theta) = 1 - \beta(\theta)$$

16.2.1 INTERPRETATION OF THE OC FUNCTION:

(i) If the true parameter is exactly θ_0 (the null hypothesis value), the OC function gives the probability of correctly not rejecting H_0, that is,

$$P(\theta) = P(Accepting\ H_0 / \theta) = P(Accepting\ H_0 / \theta = \theta_0) = 1 - \alpha$$

Since the probability of rejecting H_0 is α, the probability of accepting H_0 is 1-α.

This means that if the null hypothesis (H_0) is actually true, the probability of correctly accepting it is high.

(ii) If the true parameter is exactly θ_1 (the alternative hypothesis value), the OC function gives the probability of incorrectly accepting H_0 (Type II error, β).

$$L(\theta) = P(Accepting\ H_0 / \theta) = P(Accepting\ H_0 / \theta = \theta_1) = \beta$$

This means that if the alternative hypothesis (H_1) is true, the probability of incorrectly accepting the null hypothesis is low.

(iii) As the true parameter θ shifts from θ_0 toward θ_1, the probability of accepting H_0 decreases. This is because more evidence accumulates in favor of H_1, leading to a higher probability of rejecting H_0.

16.2.2 MATHEMATICAL EXPRESSION FOR OC FUNCTION:

If $X_1, X_2, \ldots$ is a sequence of independent observations taken from a population whose probability density / mass function is $f(X,\theta)$ which depends on a parameter θ which takes one of the two values θ_0 or θ_1, then to test a simple null hypothesis

$H_0 : \theta = \theta_0$

against a simple alternative hypothesis

$H_1 : \theta = \theta_1$

The OC function of a sequential probability ratio test is given by

$$P(\theta) = \frac{A^{h(\theta)} - 1}{A^{h(\theta)} - B^{h(\theta)}}$$

Where

- $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$ are the decision boundaries.
- $h(\theta)$ is the expected number of observations required before making a decision. We can determine its value as

$$E\left[\frac{f(X;\theta_1)}{f(X;\theta_0)}\right]^{h(\theta)} = 1$$

It has been proved that under very simple conditions on the nature of the probability density function, there exists a unique value of h(θ) such that the above condition is satisfied.

In some cases, calculating the OC function is found to be more complicated. In such cases, we may use the following formula:

$$P(\theta) = \frac{1 - B^{h(\theta)}}{1 - B^{h(\theta)} + A^{h(\theta)}}$$

where $h(\theta)$ is the expected number of samples/observations required before making a decision and given as

$$h(\theta)=\frac{(1-\beta)\log(A)+\beta\log(B)}{E[Z/\theta]}$$

To calculate the OC function, first, we have to calculate $h(\theta)$. You will study in the next session how to calculate it, therefore, we will take an example to calculate the OC function in the next session.

After understanding the OC function of the sequential probability ratio test and how we can analyse the performance of the sequential probability ratio test, let us discuss another function which measures the expected number of samples required to reach a decision in the next section.

16.3 AVERAGE SAMPLE NUMBER (ASN) FUNCTION:

The sequential probability ratio test is an efficient hypothesis testing method that does not require a fixed sample size. Instead, samples are evaluated sequentially, and the test stops as soon as enough evidence is gathered to accept or reject the null hypothesis.

One of the most important characteristics of SPRT is the Average Sample Number (ASN), which represents the expected number of samples/ observations required to reach a decision. The average sample number function denoted as $E[N/\theta]$ and is defined as follows:

The average sample number is defined as the average (expected) number of observations required to reach a decision when the true parameter is θ.

16.3.1 MATHEMATICAL EXPRESSION FOR ASN FUNCTION:

If $X_1, X_2, \ldots$ is a sequence of independent observations taken from a population whose probability density / mass function is $f(X,\theta)$ which depends on a parameter θ which takes one of the two values θ_0 or θ_1, then to test a simple null hypothesis

$$H_0:\theta=\theta_0$$

against a simple alternative hypothesis

$$H_1:\theta=\theta_1$$

The ASN function of a sequential probability ratio test is given by

$$E[N]=\frac{\{1-L(\theta)\}\log(A)+L(\theta)\log(B)}{E[Z]}$$

When the null hypothesis $(H_0:\theta=\theta_0)$ is true, then

$$P(\theta)=P(\text{Accepting } H_0/\theta)=P(\text{Accepting } H_0/\theta=\theta_0)=1-\alpha$$

When the alternative hypothesis $(H_0:\theta=\theta_1)$ is true:

$$P(\theta) = P(\textit{Accepting } H_0 / \theta) = P(\textit{Accepting } H_0 / \theta = \theta_1) = \beta$$

Therefore, we can write the expression of ASN as follows:

$$E[N/H_\theta \text{ is true}] = \frac{\alpha \log(A) + (1-\alpha)\log(B)}{E[Z/H_\theta \text{ is true}]}$$

And

$$E[N/H_1 \text{ is true}] = \frac{(1-\beta)\log(A) + \beta \log(B)}{E[Z/H_1 \text{ is true}]}$$

Let us take an example for illustration purposes.

16.4 PRACTICAL EXAMPLE:

Consider Example in 14.4 of Air Quality Index (AQI). Find ASN and OC functions for μ = 60, σ = 2, α = 0.05 and β = 0.1.

Solution: Here, it is given that

μ0 = 50, μ1 = 65, μ = 60, σ = 2, α = 0.05 and β =0.1.

The average sample function is given by

$$E[N/H_\theta \text{ is true}] = \frac{\alpha \log(A) + (1-\alpha)\log(B)}{E[Z/H_\theta \text{ is true}]}$$

To calculate ASN, we first compute these terms:

$$A = \frac{1-\beta}{\alpha} = \frac{1-0.1}{0.05} = 18$$

$$\log(A) = \log(18) = 2.89$$

$$B = \frac{\beta}{1-\alpha} = \frac{0.1}{1-0.05} = 0.105$$

$$\log(B) = \log(0.105) = -2.25$$

In Example in 14.4, we calculated the term Z as

$$Z = \frac{\mu_1 - \mu_0}{\sigma^2}\left\{X - \frac{\mu_1 + \mu_0}{2}\right\}$$

Therefore, we can calculate the expected value of Z as

$$E[Z] = \frac{(\mu_1 - \mu_0)}{\sigma^2}\left\{E[X] - \frac{\mu_1 + \mu_0}{2}\right\}$$

$$= \frac{(\mu_1 - \mu_0)}{\sigma^2}\left\{\mu - \frac{\mu_1 + \mu_0}{2}\right\} \qquad \left\{\because X \sim N(\mu, \sigma^2) \right\} \left\{\therefore E[X] = \mu\right\}$$

$$= \frac{(65-60)}{100}\left\{\mu - \frac{65-60}{2}\right\}$$

$$E[Z] = \frac{15}{200}(2\mu - 115)$$

Therefore,

$$E\left[Z/H_0 \text{ is true}\right]=\frac{15}{200}\left(2\mu_0-115\right)=\frac{15}{200}\left(2\times 50-115\right)=-1.125$$

$$E\left[Z/H_0 \text{ is true}\right]=\frac{15}{200}\left(2\mu_1-115\right)=\frac{15}{200}\left(2\times 65-115\right)=1.125$$

Thus, we can calculate ASN under H_0 and H_1 as follows:

$$E\left[N/H_0 \text{ is true}\right]=\frac{\alpha\log\left(A\right)+\left(1-\alpha\right)\log\left(B\right)}{E\left[Z/H_0 \text{ is true}\right]}$$

$$=\frac{0.05\times 2.98+0.95\times -2.25}{-1.125}$$

$$E\left[N/H_0 \text{ is true}\right]=\frac{-1.989}{-1.125}=1.768\approx 2$$

$$E\left[N/H_1 \text{ is true}\right]=\frac{\left(1-\beta\right)\log\left(A\right)+\beta\log\left(B\right)}{E\left[Z/H_1 \text{ is true}\right]}$$

$$=\frac{0.90\times 2.89+0.1\times -2.25}{-1.125}$$

$$E\left[N/H_1 \text{ is true}\right]=\frac{2.376}{1.125}=2.112\approx 3$$

Thus, on average, the SPRT requires only about 3 observations to make a decision which is much lower than a fixed-sample test.

We now calculate the OC function. We know the OC function of a sequential probability ratio test is given by

$$P\left(\theta\right)=\frac{1-B^{h(\theta)}}{1-B^{h(\theta)}+A^{h(\theta)}}$$

where $h\left(\theta\right)$ is the expected number of samples required before making a decision and given as

$$h\left(\theta\right)=\frac{\left(1-\beta\right)\log\left(A\right)+\beta\log\left(B\right)}{E\left[Z/\theta\right]}$$

Therefore, we first find $E\left[Z/\mu\right]$ as follows:

$$E\left[Z/\mu=60\right]=\frac{15}{200}\left(2\mu-115\right)$$

$$E\left[Z/\mu=60\right]=\frac{15}{200}\left(2\times 60-115\right)$$

$$E\left[Z/\mu=60\right]=0.375$$

We now calculate $h\left(\theta\right)$ as

$$h(\theta) = \frac{(1-\beta)\log(A) + \beta\log(B)}{E[Z/\theta]}$$

$$h(\theta) = \frac{0.90 \times 2.89 + 0.1 \times -2.25}{0.375}$$

$$h(\theta) = \frac{2.376}{0.375} = 6.336 \approx 7$$

Therefore, we can calculate the OC function as

$$P(\mu) = \frac{1 - B^{h(\mu)}}{1 - B^{h(\mu)} + A^{h(\mu)}}$$

$$P(\mu) = \frac{1 - (0.105)^7}{1 - (0.105)^7 + 18^7} \approx 0$$

Since P(μ = 60) is very small, it means that the probability of wrongly accepting H_0 when the average AQI has shifted to 60 is approximately zero.

16.5 RELATIONSHIP BETWEEN OC, ASN AND SPRT BOUNDARIES:

The OC and ASN functions are the primary performance measures used to characterize the efficiency of sequential probability ratio tests (SPRT).

- SPRTs use upper and lower boundaries (often denoted as **A and B,** related to the producer's risk (α) consumer risk (β) to decide whether to accept, reject, or continue sampling.
- The specific location of these boundaries, determined by the desired risks (α and β) directly impacts the shape of both the OC and ASN curves.
- The ASN for an SPRT is typically much lower than for a fixed sample size test, especially when the true quality is either very good or very bad, requiring only a few samples to make a decision. The ASN function generally peaks near the "action threshold" or indifference quality level, where discrimination is most difficult and thus more samples are needed to reach a conclusion.

16.6 APPLICATIONS:

OC and ASN functions have wide applications beyond manufacturing, including:

- **Ecology and Pest Management**: Determining sampling plans for monitoring pest populations in fields to decide whether to apply treatment.
- **Clinical Trials**: Designing sequential tests for comparing the effectiveness of new drugs, balancing the need for reliable results with minimizing the number of patients or duration of the trial.
- **Auditing**: Deciding the optimal sample size for financial audits to ensure a certain level of confidence in the financial records
- **Industrial Quality Control**: Acceptance sampling, production inspection.
- **Reliability Testing**: Failure detection, lifetime testing.
- **Surveillance and Detection**: Signal processing, intrusion detection.

- **Econometrics & Finance**: Sequential decision making in markets.
- **Machine Learning**: Online learning, adaptive testing.

16.7 SUMMARY:

The OC and ASN functions are fundamental tools in statistics for designing and evaluating sampling plans and hypothesis tests. The OC function quantifies the probability of making correct or incorrect decisions for varying population quality levels, while the ASN function quantifies the average amount of sampling (cost) required to reach a decision. Together, they allow practitioners to balance the risks of wrong decisions with the costs of inspection.

- The **OC Function** describes the probability of accepting H_0 for different parameter values.
- The **ASN Function** gives expected sample size needed by the SPRT.
- Both functions are essential to evaluate the performance of the SPRT.
- OC and ASN depend on decision boundaries **A and B,** which in turn depend on error probabilities.
- SPRT provides extremely efficient testing compared to fixed-sample procedures.

16.8 KEY WORDS:

The most important keywords for this topic are:

- Acceptance Sampling
- Acceptable Quality Level (AQL)
- Average Sample Number (ASN)
- Consumer's Risk (β)
- Lot Tolerance Percent Defective (LTPD)
- Operating Characteristic (OC) Curve/Function
- Producer's Risk (α)
- Sequential Probability Ratio Test (SPRT)

16.9 SELF-ASSESSMENT QUESTIONS:

1. Define the Operating Characteristic (OC) function of an SPRT.
2. What is the role of ASN in evaluating a sequential test?
3. Derive the approximate relationship between ASN and OC.
4. Explain how decision boundaries affect OC and ASN.
5. Compute A and B for $\alpha = 0.01, \beta = 0.05$
6. Let X have the distribution $f(x,\theta) = \theta^x (1-\theta)^{1-x}; x = 0,1; 0 < \theta < 1$ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, construct S.P.R.T. and obtain it's A.S.N. and O.C. function.
7. For a Bernoulli SPRT, derive $Z = \ln \frac{f_1(X)}{f_0(X)}$.
8. Why is the ASN larger near the "indifference" region?
9. Explain the applications of OC and ASN functions in sequential testing.

16.10 SUGGESTED READINGS:

- Goon, A.M., Gupta, M.K., and Dasgupta, B.: *Fundamentals of Statistics*.
- Miller, Irwin and Miller, Marylees: *John E. Freund's Mathematical Statistics with Applications*.
- Mood, A.M., Graybill, F.A., and Boes, D.C.: *Introduction to the Theory of Statistics*.
- Johnson, R.A., and Bhattacharya, G.K.: *Statistics-Principles and Methods*.
- Wald, Abraham: *Sequential Analysis* (a classic text for SPRT).
- Mathematical Statistics- Parimal Mukopadhyay(1996), New Central Book Agency (P)Ltd., Calcutta
- Statistical Inference by H.C, Saxena & Surendran
- An outline of Statistical Theory, vol.2 by A.M. Goon and B. Das Gupta
- An Introduction to probability and Mathematical Statistics by V.K. Rohatgi

Dr. K. Kalyani

LESSON -17

APPLICATIONS OF BINOMIAL, POISSON, NORMAL DISTRIBUTIONS & SEQUENTIAL TESTING EFFICIENCY

OBJECTIVES:

By the end of this lesson, students will be able to:

- ❖ Understand how the Sequential Probability Ratio Test (SPRT) is applied to common statistical distributions such as Binomial**,** Poisson and Normal**.**
- ❖ Derive likelihood ratios and decision boundaries for each distribution under simple hypotheses.
- ❖ Compute stopping boundaries and implement practical sequential testing procedures.
- ❖ Evaluate the efficiency of sequential tests relative to fixed-sample-size tests.
- ❖ Apply SPRT principles in real-life scenarios where data arrive sequentially.
- ❖ Interpret OC and ASN functions in the context of these distributions.
- ❖ Identify advantages and limitations of sequential tests across distributions.

STRUCTURE:

17.1 INTRODUCTION:

Sequential analysis deals with statistical procedures in which the sample size is not fixed in advance**,** but is determined by the data as they are observed. Among all sequential

procedures, the Sequential Probability Ratio Test (SPRT) proposed by Abraham Wald is one of the most important and widely used methods for testing statistical hypotheses.

In a classical (fixed-sample) test, the researcher collects a predetermined number of observations and then makes a decision. However, in many practical situations—such as industrial inspection, clinical trials, quality control, or communication systems—data arrive one at a time**,** and it is desirable to reach a conclusion as early as possible**.** Collecting unnecessary samples increases cost, time, and risk. SPRT provides an optimal framework for such real-time decision making.

The SPRT is based on evaluating the likelihood ratio after each observation and comparing it with two decision boundaries. Depending on where the likelihood ratio falls, the test may:

- accept the null hypothesis H_0,
- accept the alternative H_1, or
- continue sampling.

This procedure is repeated until enough evidence is accumulated in favour of either hypothesis. An important feature of SPRT is that, for given Type I and Type II error probabilities (α and β), it is the most efficient test among all sequential tests and typically requires fewer observations than fixed-sample procedures.

Because many real-life data sets follow binomial**,** Poisson**,** or normal distributions, the application of SPRT to these models is especially important:

- Binomial distribution arises in situations involving success–failure data, such as defect detection or medical treatment response.
- Poisson distribution describes counts of rare events, such as accidents, machine breakdowns, or call arrivals.
- Normal distribution models continuous measurements like weight, length, pressure, or temperature.

In each of these cases, the likelihood ratio and decision process take a specific mathematical form, allowing SPRT to be implemented efficiently. The efficiency of sequential tests is studied using measures such as the Average Sample Number (ASN) and the Operating Characteristic (OC) function**,** which help quantify the reduction in sampling effort.

Thus, the study of SPRT applications to binomial, Poisson, and normal models, along with its efficiency, forms a crucial part of sequential analysis and plays a significant role in modern statistical decision making.

17.2 SPRT FOR BINOMIAL DISTRIBUTION:

Let X has the distribution $f(x,\theta)=\theta^x(1-\theta)^{1-x}; x=0,1 \;\; 0<\theta<1$ for testing $H_0:\theta=\theta_0$ *against* $H_1:\theta=\theta_1$, construct SPRT and obtain its ASN and OC function.

$$\lambda_m = \frac{L(x_1, x_{2.}, ..., x_m / H_1)}{L(x_1, x_{2.}, ..., x_m / H_0)}$$

$$\lambda_m = \left[\theta_1^{\sum x_1}(1-\theta_1)^{m-\sum x_i}\right] \Big/ \left[\theta_0^{\sum x_1}(1-\theta_0)^{m-\sum x_i}\right]$$

$$\lambda_m = \left(\frac{\theta_1}{\theta_0}\right)^{\sum x_1} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{m-\sum x_i}$$

$$l\log \lambda_m = \sum x_i \log\left(\frac{\theta_1}{\theta_0}\right) + \left(m - \sum x_i\right)\log\left(\frac{1-\theta_1}{1-\theta_0}\right)$$

$$\log \lambda_m = \sum x_i \log\left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right] + m \log\left(\frac{1-\theta_1}{1-\theta_0}\right)$$

Hence, SPRT for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$

(i) Accept H_0 if $\log \lambda_m \le \log\left(\frac{\beta}{1-\alpha}\right) = b,$

(ii) If $\sum x_i \le \frac{b - m \log\left[(1-\theta_1)/(1-\theta_0)\right]}{\log\left[\theta_1(1-\theta_0)/\theta_0(1-\theta_1)\right]} = a_m$

(iii) Reject H_0 if $\log \lambda_m \ge \log\left(\frac{1-\beta}{\alpha}\right) = a,$

If $\sum x_i \ge \frac{a - m \log\left[(1-\theta_1)/(1-\theta_0)\right]}{\log\left[\theta_1(1-\theta_0)/\theta_0(1-\theta_1)\right]} = r_m$

Continue sampling if $b < \log \lambda_m < a \Rightarrow a_m < \sum x_i < r_m$

OC function

OC function is given by

$$L(\theta) = \frac{A^{h(\theta)} - 1}{A^{h(\theta)} - B^{h(\theta)}}$$

Where for each value of θ, $h(\theta) \neq 0$ is to be determined such that

$$E\left[\frac{f(x,\theta_1)}{f(x,\theta_0)}\right]^{h(\theta)} = 1$$

$$\sum_{x=0}^{1}\left[\frac{f(x,\theta_1)}{f(x,\theta_0)}\right]^{h(\theta)} f(x,\theta) = 1$$

$$\sum_{x=0}^{1}\left[\left(\frac{\theta_1}{\theta_0}\right)^x\left(\frac{1-\theta_1}{1-\theta_0}\right)^{1-x}\right]^{h(\theta)}\theta(1-\theta)^{1-x}=1$$

$$\left[\left(\frac{1-\theta_1}{1-\theta_0}\right)^{h(\theta)}.(1-\theta)+\left(\frac{\theta_1}{\theta_0}\right)^{h(\theta)}\right]\theta=1$$

This equation for $h=h(\theta)$ is very tedious from practical point of view, instead of solving for h we regard h as a parameter and solve it for θ. Thus giving

$$\theta\left[\left(\frac{\theta_1}{\theta_0}\right)^{h(\theta)}-\left(\frac{1-\theta_1}{1-\theta_0}\right)^{h(\theta)}\right]=1-\left(\frac{1-\theta_1}{1-\theta_0}\right)^{h(\theta)}$$

$$\Rightarrow\theta=\frac{1-\left[(1-\theta_1)/(1-\theta_0)\right]^{h(\theta)}}{(\theta_1/\theta_0)^{h(\theta)}-\left[(1-\theta_1)/(1-\theta_0)\right]^{h(\theta)}}=\theta(h)$$

$$\therefore L(\theta)=\frac{\left[(1-\beta)/\alpha\right]^h-1}{\left[(1-\beta)/\alpha\right]^h-\left[\beta/(1-\alpha)\right]^h}=L(\theta,h)$$

Various points on the OC curve are obtained by assigning arbitrary values to 'h' and computing the corresponding values of θ and $L(\theta)$.

ASN Function

$$Z=\log\left[\frac{f(x,\theta_1)}{f(x,\theta_0)}\right];\ \text{A}=\frac{1-\beta}{\alpha}\ \text{B}=\frac{\beta}{1-\alpha}$$

$$E(Z)=\sum_{x=0}^{1}\log\left[\frac{f(x,\theta_1)}{f(x,\theta_0)}\right].f(x,\theta)$$

$$=\sum_{x=0}^{1}\log\left[\left(\frac{\theta_1}{\theta_0}\right)^x\left(\frac{1-\theta_1}{1-\theta_0}\right)^{1-x}\right]^{h(\theta)}\theta^x(1-\theta)^{1-x}$$

$$=(1-\theta)\log\left(\frac{1-\theta_1}{1-\theta_0}\right)+\theta\ \log\left(\frac{\theta_1}{\theta_0}\right)$$

$$E(Z)=\theta\log\left[\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right]+\log\left(\frac{1-\theta_1}{1-\theta_0}\right)$$

ASN is given by

$$E(n)=\frac{L(\theta)\log B+\left[1-L(\theta)\right]\log A}{E(Z)}$$

Substituting the values of $E(Z)$ and $L(\theta)$, we get ASN function.

17.3 SPRT FOR POISSON DISTRIBUTION:

Example 1: First of all, we formulate the null hypothesis (H_0) and alternative hypothesis (H_1) as follows:

$H_0 : \lambda = \lambda_0 = 2$ (average defect rate per hour is 2)

$H_1 : \lambda = \lambda_1 = 5$ (average defect rate per hour is 5)

Since the number of defective parts per hour follows the Poisson distribution, therefore, the probability mass function of the Poisson distribution with parameter λ is given as follows:

$$P[X = x] = \frac{e^{-\lambda}\lambda^{x}}{x!}; X = 0,1,... \& \lambda > 0$$

Solution: We now compute the probability mass function under the null hypothesis (H_0) and alternative hypothesis (H_1) as follows:

$$P[X = x_i; \lambda_0] = \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}$$

Similarly, $P[X = x_i; \lambda_1] = \frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!}$

Therefore, we can compute $Z_i = \log\left\{\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}\right\}$ as follows:

$$Z_i = \log\left\{\frac{f(X_i;\theta_1)}{f(X_i;\theta_0)}\right\}$$

$$= \log\left\{\frac{\frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!}}{\frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}}\right\}$$

$$= \log\left\{\left(\frac{\lambda_1}{\lambda_0}\right)^{x_i} e^{-(\lambda_1-\lambda_0)}\right\}$$

$$Z_i = -(\lambda_1 - \lambda_0) + X_i \log\left\{\left(\frac{\lambda_1}{\lambda_0}\right)^{x_i}\right\}$$

Hence,

$$\sum_{i=1}^{m} Z_i = -m(\lambda_1 - \lambda_0) + \log\left(\frac{\lambda_1}{\lambda_0}\right)^{x_i} \sum_{i=1}^{m} X_i$$

Hence, the SPRT for testing $H_0 : \lambda = \lambda_0 = 2$ against $H_1 : \lambda = \lambda_1 = 5$ is given as

(i) Reject H_0 if

$$\sum_{i=1}^{m} Z_i \geq \log\left(\frac{1-\beta}{\alpha}\right)$$

$$-m(\lambda_1-\lambda_0)+\log\left(\frac{\lambda_1}{\lambda_0}\right)^{x_i}\sum_{i=1}^{m} X_i \geq \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\sum_{i=1}^{m} X_i \geq \frac{1}{\log\left(\frac{\lambda_1}{\lambda_0}\right)^{x_i}}\left\{\log\left(\frac{1-\beta}{\alpha}\right)+m(\lambda_1-\lambda_0)\right\}$$

(ii) Accept H_0 if

$$\sum_{i=1}^{m} Z_i \leq \log\left(\frac{\beta}{1-\alpha}\right)$$

$$\sum_{i=1}^{m} X_i \leq \frac{1}{\log\left(\frac{\lambda_1}{\lambda_0}\right)^{x_i}}\left\{\log\left(\frac{\beta}{1-\alpha}\right)+m(\lambda_1-\lambda_0)\right\}$$

(iii) Continue taking additional observations as long as

$$\log\left(\frac{\beta}{1-\alpha}\right) < \sum_{i=1}^{m} Z_i < \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\log\left(\frac{\beta}{1-\alpha}\right) < -m(\lambda_1-\lambda_0)+\log\left(\frac{\lambda_1}{\lambda_0}\right)^{x_i}\sum_{i=1}^{m} X_i < \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\frac{1}{\log\left(\frac{\lambda_1}{\lambda_0}\right)^{x_i}}\left\{\log\left(\frac{\beta}{1-\alpha}\right)+m(\lambda_1-\lambda_0)\right\} < \sum_{i=1}^{m} X_i < \frac{1}{\log\left(\frac{\lambda_1}{\lambda_0}\right)^{x_i}}\left\{\log\left(\frac{1-\beta}{\alpha}\right)+m(\lambda_1-\lambda_0)\right\}$$

Here, it is given that

$\lambda_0 = 2, \lambda_1 = 5, \alpha = 0.1$ and β=0.1

Therefore, we compute

$$\log\left(\frac{\lambda_1}{\lambda_0}\right) = \log\left(\frac{5}{2}\right) = 0.92$$

$$A = \frac{1-\beta}{\alpha} = \frac{0.9}{0.1} = 9$$

Similarly,

$$\log(A) = \log(9) = 2.20$$

$$B = \frac{\beta}{1-\alpha} = \frac{0.1}{1-0.1} = 0.11$$

$$\log(B) = \log(0.11) = -2.21$$

Putting these values in the SPRT test, we get the SPRT for testing $H_0 : \lambda = \lambda_0 = 2$ against $H_1 : \lambda = \lambda_1 = 5$ is given as

(i) Reject H_0 if

$$\sum_{i=1}^{m} X_i \geq 1.09(2.20 + 3m)$$

(ii) Accept H_0 if

$$\sum_{i=1}^{m} X_i \leq 1.09(3m - 2.21)$$

(iii) Continue taking additional observations as long as

$$1.09(3m - 2.21) < \sum_{i=1}^{m} X_i < 1.09(2.20 + 3m)$$

Example 2: Here, it given that

$\lambda_0 = 2, \lambda_1 = 5, \alpha = 0.05$ and β=0.1

In Example 1, we calculated

$$\log\left(\frac{\lambda_1}{\lambda_0}\right) = 0.92, A = 9, \log(A) = 2.20, B = 0.11, \log(B) = -2.21$$

We can calculate value of Z as follows:

$$Z = -(\lambda_1 - \lambda_0) + X \log\left(\frac{\lambda_1}{\lambda_0}\right) = -3 + 0.92X$$

Thus,

$$E[Z/H_0 \text{ is true}] = -3 + 0.92E[X]$$

$$= -3 + 0.92 * \lambda_0 \quad \begin{Bmatrix} \because X \sim Poiss(\lambda) \\ \therefore E[X] = \lambda \end{Bmatrix}$$

$$= -3 + 0.92 * 2 = -1.168$$

$$E[Z/H_1 \text{ is true}] = -3 + 0.92E[X]$$

$$= -3 + 0.92 * \lambda_1$$

$$= -3 + 0.92 * 5 = 1.58$$

Therefore, the average sample umber is given by

$$E\left[N/H_0 \text{ is true}\right] = \frac{\alpha\log(A)+(1-\alpha)\log(B)}{E\left[Z/H_0 \text{ is true}\right]}$$

$$= \frac{0.1\times 2.20 + 0.9\times -2.21}{-1.168}$$

$$E\left[N/H_0 \text{ is true}\right] = \frac{-1.769}{-1.168} = 1.51 \approx 2$$

$$E\left[N/H_1 \text{ is true}\right] = \frac{(1-\beta)\log(A)+\beta\log(B)}{E\left[Z/H_1 \text{ is true}\right]}$$

$$= \frac{0.90\times 2.20 + 0.1\times -2.21}{-1.58}$$

$$E\left[N/H_1 \text{ is true}\right] = \frac{1.759}{1.58} = 1.11$$

Thus, on average, the SPRT requires only about 2 observations to make a decision which is much lower than a fixed-sample test.

We now calculate the OC function. We know the OC function of a sequential probability ratio test is given by

$$P(\theta) = \frac{1-B^{h(\theta)}}{1-B^{h(\theta)}+A^{h(\theta)}}$$

where $h(\theta)$ is the expected number of samples required before making a decision and given as $h(\theta) = \dfrac{(1-\beta)\log(A)+\beta\log(B)}{E\left[Z/\theta\right]}$

Therefore,

$$E\left[Z/\lambda = 4\right] = -3 + 0.916E\left[X\right]$$

$$= -3 + 0.916 * 4 = 0.664$$

We calculate $h(\theta)$ as

$$h(\theta) = \frac{(1-\beta)\log(A)+\beta\log(B)}{E\left[Z/\theta\right]}$$

$$= \frac{0.90\times 2.20 + 0.1\times -2.21}{0.664}$$

$$= \frac{1.759}{0.664} = 2.65$$

Therefore, we can calculate the OC function as

$$P(\lambda) = \frac{1 - B^{h(\lambda)}}{1 - B^{h(\lambda)} + A^{h(\lambda)}}$$

$$P(\mu) = \frac{1 - (0.11)^{2.65}}{1 - (0.11)^{2.65} + (2.20)^{2.65}}$$

$$= \frac{0.997}{9.077} = 0.11$$

Since L(4) is small, it means that the probability of wrongly accepting H_0 when the defect rate increases to 4 per hour is approximately 11%.

17.4 SPRT FOR NORMAL DISTRIBUTION:

We set up the hypothesis for the variance as follows:

$H_0 : \sigma^2 = \sigma_0^2 = 1$ (The machine is working fine, variance remains at 1)

$H_0 : \sigma^2 = \sigma_0^2 = 2$ (The machine is malfunctioning, variance increases to 2)

The deviation from the expected whole size follows the normal distribution with mean 0 and variance σ^2, therefore, we can write the probability density function of the normal distribution as follows:

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

We now compute the pdf under the null hypothesis (H_0) and alternative hypothesis (H_1) as follows:

$$f(X;\sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{x^2}{2\sigma_0^2}}$$

Similarly,

$$f(X;\sigma_1^2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{x^2}{2\sigma_1^2}}$$

Therefore, we can compute $Z_i = \log\left\{\frac{f(X_i;\sigma_1^2)}{f(X_i;\sigma_0^2)}\right\}$ as follows:

$$Z_i = \log\left\{\frac{f\left(X_i;\sigma_1^2\right)}{f\left(X_i;\sigma_0^2\right)}\right\} = \log\left\{\frac{\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{x^2}{2\sigma_1^2}}}{\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{x^2}{2\sigma_0^2}}}\right\}$$

$$= \log\left\{\left(\frac{\sigma_0^2}{\sigma_1^2}\right)^{1/2} e^{\frac{x^2}{2}\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)}\right\}$$

We now try to simplify it as follows:

$$Z_i = \frac{1}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) + \frac{x_i^2}{2}\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)$$

$$= \frac{1}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) + \frac{x_i^2}{2}\left(\frac{\sigma_1^2-\sigma_0^2}{\sigma_0^2\sigma_1^2}\right)$$

$$= \frac{1}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) + x_i^2\left(\frac{\sigma_1^2-\sigma_0^2}{2\sigma_0^2\sigma_1^2}\right)$$

$$\sum_{i=1}^{m} Z_i = \frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) + \sum_{i=1}^{m} x_i^2\left(\frac{\sigma_1^2-\sigma_0^2}{2\sigma_0^2\sigma_1^2}\right)$$

Hence, the SPRT for testing $H_0 : \sigma^2 = \sigma_0^2 = 1$ against $H_0 : \sigma^2 = \sigma_0^2 = 2$ is given as

(i) Reject H_0 if

$$\sum_{i=1}^{m} Z_i \geq \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) + \sum_{i=1}^{m} x_i^2\left(\frac{\sigma_1^2-\sigma_0^2}{2\sigma_0^2\sigma_1^2}\right) \geq \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\sum_{i=1}^{m} x_i^2 \geq \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left\{\log\left(\frac{1-\beta}{\alpha}\right) - \frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right\}$$

(ii) Accept H_0 if

$$\sum_{i=1}^{m} Z_i \leq \log\left(\frac{\beta}{1-\alpha}\right)$$

$$\sum_{i=1}^{m} x_i^2 \leq \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left\{\log\left(\frac{\beta}{1-\alpha}\right) - \frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right\}$$

(iii) Continue taking additional observations as long as

$$\log\left(\frac{\beta}{1-\alpha}\right) < \sum_{i=1}^{m} Z_i < \log\left(\frac{1-\beta}{\alpha}\right)$$

$$\frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left\{\log\left(\frac{\beta}{1-\alpha}\right)-\frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right\} \le \sum_{i=1}^{m} x_i^2 \le \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left\{\log\left(\frac{1-\beta}{\alpha}\right)-\frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right\}$$

Here, it is given that

$\sigma_0^2 = 1, \sigma_1^2 = 2, \alpha = 0.05$ and β=0.10

Therefore,

$$\log\left(\frac{\beta}{1-\alpha}\right) = \log\left(\frac{0.1}{1-0.05}\right) = \log(0.105) = -2.25$$

$$\log\left(\frac{1-\beta}{\alpha}\right) = \log\left(\frac{1-0.1}{0.05}\right) = \log(18) = 2.89$$

$$\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) = \log\left(\frac{1}{2}\right) = -0.69$$

$$\frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2} = \frac{2\times1\times2}{2-1} = 4$$

(i) Reject H_0 if

$$\sum_{i=1}^{m} x_i^2 \ge \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left\{\log\left(\frac{1-\beta}{\alpha}\right)-\frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right\} = 4\left\{2.89-\frac{m}{2}\times-0.69\right\}$$

$$\sum_{i=1}^{m} x_i^2 \ge 4(2.89+0.345m)$$

(ii) Accept H_0 if

$$\sum_{i=1}^{m} x_i^2 \le \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left\{\log\left(\frac{\beta}{1-\alpha}\right)-\frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right\} = 4\{-2.25+0.345m\}$$

(iii) Continue taking additional observations as long as

$$\frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left\{\log\left(\frac{\beta}{1-\alpha}\right)-\frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right\} \le \sum_{i=1}^{m} x_i^2 \le \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\left\{\log\left(\frac{1-\beta}{\alpha}\right)-\frac{m}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right\}$$

$$4(-2.25+0.345) \le \sum_{i=1}^{m} x_i^2 \le 4(2.89+0.345m)$$

We now show the step-by-step decision using SPRT.

We have

$$Z = \frac{1}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) + X^2\left(\frac{\sigma_1^2-\sigma_0^2}{2\sigma_0^2\sigma_1^2}\right)$$

$$Z=\frac{1}{2}\times -0.69+\frac{X^2}{4}$$
$$=-0.345+\frac{X^2}{4}$$

We calculate $\sum_{i=1}^{m} Z_i$ at each step and compare it with the log(A) and log(B) and take the decision about the null hypothesis H_0 as shown in the following table:

Sample	X	X^2	$Z=-0.345+\frac{X^2}{4}$	Cumulative Sum $\sum_{i=1}^{m} Z_i$	Decision
1	0.5	0.25	-0.283	-0.283	Continue
2	0.6	0.36	1.095	0.813	Continue
3	0.7	0.49	1.615	2.428	Continue
4	0.8	0.64	2.215	4.643	Continue

Hence, the SPRT method efficiently detected the defect rate increases with only 4 observations/samples instead of using a fixed-sample test (which may require 30 or more than 30 observations).

We now compute the average sample number.

$$E\left[Z/H_0 \text{ is true}\right]=-0.345+\frac{1}{4}E\left[X^2\right]\begin{Bmatrix}\because X \square N\left(0,\sigma^2\right)\\ \therefore E\left[X^2\right]=\sigma^2\end{Bmatrix}$$
$$=-0.345+\frac{1}{4}\sigma_0^2$$
$$=-0.345+\frac{1}{4}\times 1$$
$$=-0.345+0.25$$
$$E\left[Z/H_0 \text{ is true}\right]=-0.095$$

$$E\left[Z/H_1 \text{ is true}\right]=-0.345+\frac{1}{4}\sigma_1^2$$
$$=-0.345+\frac{2}{4}$$
$$=-0.345+0.50$$
$$E\left[Z/H_1 \text{ is true}\right]=0.155$$

Therefore, the average sample function is given by

$$E\left[N/H_0 \text{ is true}\right]=\frac{\alpha\log\left(A\right)+\left(1-\alpha\right)\log\left(B\right)}{E\left[Z_1/H_0 \text{ is true}\right]}$$

$$E\left[N/H_0 \text{ is true}\right]=\frac{0.05\times 2.98+0.95\times -2.25}{-0.095}$$

$$E\left[N/H_0 \text{ is true}\right]=\frac{-1.989}{-0.095}$$

$$E\left[N/H_0 \text{ is true}\right]=20.94\approx 21$$

$$E\left[N/H_1 \text{ is true}\right]=\frac{\left(1-\beta\right)\log\left(A\right)+\beta\log\left(B\right)}{E\left[Z_1/H_1 \text{ is true}\right]}$$

$$E\left[N/H_1 \text{ is true}\right]=\frac{0.90\times 2.89+0.1\times 2.25}{0.155}$$

$$E\left[N/H_1 \text{ is true}\right]=\frac{2.826}{0.155}$$

$$E\left[N/H_1 \text{ is true}\right]=18.23\approx 19$$

We now calculate the OC function. We know the OC function of a sequential probability ratio test is given by

$$P\left(\theta\right)=\frac{1-B^{h(\theta)}}{1-B^{h(\theta)}+A^{h(\theta)}}$$

Where $h\left(\theta\right)$ is the expected number of samples required before making a decision and given as

$$h\left(\theta\right)=\frac{\left(1-\beta\right)\log\left(A\right)+\beta\log\left(B\right)}{E\left[Z/\theta\right]}$$

Therefore,

$$Z==0.345+\frac{1}{4}X^2$$

$$E\left[Z/\sigma^2=1.5\right]=0.345+\frac{1}{4}X^2$$

$$E\left[Z/\sigma^2=1.5\right]=0.345+\frac{1}{4}\sigma^2$$

$$E\left[Z/\sigma^2=1.5\right]=0.345+\frac{1}{4}\sigma^2$$

$$E\left[Z/\sigma^2=1.5\right]=-0.345+\frac{1.5}{4}$$

$$E\left[Z/\sigma^2=1.5\right]=0.03$$

We find the average sample number $h(\theta)$ for σ^2=1.5 as follows:

$$E\left[N/\sigma^2=1.5\right]=\frac{(1-\beta)\log(A)+\beta\log(B)}{E\left[Z_1/\sigma^2=1.5\right]}$$

$$E\left[N/\sigma^2=1.5\right]=\frac{0.90\times 2.89+0.1\times 2.25}{0.03}$$

$$E\left[N/\sigma^2=1.5\right]=\frac{2.826}{0.03}$$

$$E\left[N/\sigma^2=1.5\right]=94.2\approx 95$$

$$P(\sigma^2)=\frac{1-B^{h(\sigma^2)}}{1-B^{h(\sigma^2)}+A^{h(\sigma^2)}}$$

$$P(\sigma^2)=\frac{1-(0.105)^{95}}{1-(0.105)^{95}+(2.89)^{95}}\approx 0$$

Since P(1.5) is very small, it means that the probability of wrongly accepting H_0 when the average AQI has shifted to 60 is approximately zero.

17.5 EFFICIENCY OF A SEQUENTIAL TEST:

Sequential tests are designed to reach a decision using, on average, fewer observations than fixed-sample tests while maintaining the same error probabilities. The concept of efficiency in sequential testing is primarily measured using the Average Sample Number (ASN)**.** Wald's Sequential Probability Ratio Test (SPRT) is known to be the *most efficient* sequential test in a precise optimality sense.

In general many different tests may be derived for given $\alpha \& \beta, \theta_0 \& \theta_1$. There is no point in comparing their power for given sample numbers because they are arranged. So as to have the same β - errors. We may how ever define efficiency in terms of sample size of ASN.

The test with the smaller ASN may reasonably be said to be more efficient following wald (1947) we shall prove that when end effects are negligible the SPRT is most efficient test. More precisely if 'S' is a SPRT and S is some other test based on the sum of the logarithms of identically distributed variables.

$$E_i(n/s)\geq E_i(n/s');\ \text{i} = 0, 1.....(1)$$

Where E_i denotes the expected value of 'm' on hypothesis.

Note, first of all that if 'u' is any random variable $u-E(u)$ is the value measured from the mean.

and $Exp\{u-E(u)\}\geq 1+\{u-E(u)\}$

on taking expectation we have,

$$E\left[exp\left\{u-E(u)\right\}\right]\geq 1.....(2)$$

Which gives,

$$\mathrm{E}\left(\exp.u\right)\geq \exp\left\{E(u)\right\}.....(3)$$

We also have for any closed sequential test based on the sum of type Z_n.

$$E_i\left(n/s\right)=\frac{E_i\left(\log\left(L_n(S)\right)\right)}{E_i(Z)}.....(4)$$

If E^* denotes the conditional expectation when H_0 is true and E^{**} the conditional expectation when H_1 is true, neglecting end effects.

$$\mathrm{E}^*\left(L_n/S\right)=\frac{\beta}{1-\alpha}.....(5)$$

and similarly

$$\mathrm{E}^{**}\left(L_n/S\right)=\frac{1-\beta}{\alpha}.....(6)$$

Hence,

$$\mathrm{E}_0\left(n/s\right)=\frac{1}{E_0(Z)}\left\{(1-\alpha)\mathrm{E}^*\left(\log L_n/S\right)+\alpha.E^{**}\left(\log L_n/S\right)\right\}.....(7)$$

In virtue of (3), (5), (6) equations we have

$$\mathrm{E}_0(Z)\leq 0$$

$$\mathrm{E}_0\left(n/s\right)\geq \frac{1}{E_0(Z)}\left\{(1-\alpha)\log\frac{\beta}{1-\alpha}+\alpha.\log\frac{1-\beta}{\alpha}\right\}.....(8)$$

and interchanging H_0 & H_1. α & β in equation (8) given

$$\mathrm{E}_1\left(n/s\right)\geq \frac{1}{E_1(Z)}\left\{\beta\log\frac{\beta}{1-\alpha}+(1-\beta)\log\frac{1-\beta}{\alpha}\right\}.....(9)$$

When $\mathrm{S}=\mathrm{S}^1$ these inequalities are replaced by equalities

Efficiency Advantages:

- **Reduced Sample Size:** Generally requires significantly fewer samples than traditional fixed-sample tests (like Neyman-Pearson), saving time and cost.
- **Dynamic Decisions:** Allows for timely decisions in dynamic environments (e.g., manufacturing, medicine, AI testing).

17.6 APPLICATIONS:

Applications of Binomial Distribution

The binomial distribution models the probability of a specific number of successes in a fixed number of independent trials, where each trial has only two outcomes (success/failure).

- **Quality Control:** Manufacturers use it to determine the probability of a certain number of defective items in a batch of products, ensuring quality standards are met.
- **Medicine:** It helps assess a new drug's effectiveness by modeling the probability of a patient being cured or experiencing side effects.
- **Finance/Insurance:** Banks and insurance companies use it for risk assessment, such as modeling the number of loan defaults or fraudulent transactions within a given period.
- **Market Research:** Businesses use "yes/no" surveys to predict consumer preferences or behaviors for new products.
- **Computer Networking:** It can determine the probability of a certain number of users transmitting data simultaneously on a network.

Applications of Poisson Distribution

The Poisson distribution models the number of times an event occurs within a fixed interval of time or space, given a constant average rate.

- **Traffic Analysis:** City planners and insurance companies use it to predict the number of cars arriving at a traffic light per hour or the number of accidents per month to inform road safety measures and insurance pricing.
- **Healthcare:** Hospitals use it to model the number of patients arriving at an emergency room per day, helping them manage staffing levels and resources.
- **Telecommunications:** Call centers use it to model the number of calls received per minute to optimize staffing and ensure minimal customer wait times.
- **Ecology/Astronomy:** It can estimate the number of trees of a certain species in a forest area or the number of meteorites striking the Earth per year.
- **Manufacturing:** It helps analyze the number of defects (e.g., typographical errors in a book or flaws in a fabric roll) per unit to maintain quality.

Applications of Normal Distribution

The normal distribution (or "bell curve") is a continuous probability distribution that is fundamental in statistics and finance due to its symmetry around the mean. It describes many natural phenomena and is key to the Central Limit Theorem.

- **Human Characteristics:** It is used to model physical attributes like height, weight, and shoe size in a population.
- **Testing and Scores:** Standardized test scores (like the ACT or GMAT) and IQ scores are designed to follow a normal distribution, allowing for easy comparison of individual performance relative to the average.
- **Quality Control:** Manufacturers measure product weights or dimensions, using the normal distribution to determine if variations are due to random chance or a production issue that needs addressing.
- **Finance:** It is used to model stock prices and asset returns, although real-world data often shows "fat tails" (more extreme movements than the normal distribution predicts).

- **Engineering/Architecture:** The distribution of human height is used to determine optimal design parameters, such as the standard height of doors, to accommodate most people comfortably.

Efficiency of a Sequential Test

Sequential analysis is a statistical method that analyzes data as it is collected, allowing for conclusions to be reached before the entire predetermined sample size is gathered. The main advantage is its **efficiency**, as it can lead to significant savings in time and resources.

- **Clinical Trials:** This is a key application. If interim analysis of a new medication's data shows it is overwhelmingly effective or unexpectedly harmful, the trial can be stopped early. This brings beneficial treatments to market sooner or prevents further patient exposure to an unsafe treatment more efficiently than a traditional fixed-sample trial.
- **A/B Testing (Website/Marketing):** In digital marketing, companies test different website layouts or ad designs. Sequential testing allows them to "peek" at the results continuously and stop the test as soon as a clear "winner" emerges, saving time and allowing for faster data-driven decisions.
- **Quality Control in Manufacturing:** Instead of testing a large, fixed batch of ammunition (as in its historical origins), sequential testing allows for item-by-item inspection with the possibility of stopping the process early if quality thresholds are consistently met or missed, making the inspection process more efficient.
- **Cyber security:** By continuously analyzing network traffic, sequential methods can detect anomalies indicative of an intrusion in real-time, enabling prompt mitigation of potential damage.

17.7 SUMMARY:

This lesson focuses on the application of Wald's Sequential Probability Ratio Test (SPRT) to three major probability models-Binomial**,** Poisson and Normal distributions—commonly encountered in statistical inference and quality control.

The lesson begins with a review of key probability distributions and the foundational concepts required for constructing likelihood ratios. For each distribution, the SPRT is formulated by deriving the sequential likelihood ratio, defining decision boundaries based on pre-specified Type I and Type II error probabilities, and developing stopping rules. Worked-out examples are included for the Binomial, Poisson, and Normal cases to illustrate computation of the sequential likelihood ratio after each observation and the decision-making process.

Further, the concept of efficiency of sequential tests is discussed, emphasizing how SPRT typically requires a smaller Average Sample Number (ASN) compared to fixed-sample tests while maintaining specified error rates. Practical applications in quality control, reliability testing, industrial production, medical trials, and real-time monitoring systems are highlighted.

17.8 KEY WORDS:

- Sequential Probability Ratio Test (SPRT)
- Sequential Testing
- Likelihood Ratio
- Decision Boundaries (A and B)
- Type I Error (α)
- Type II Error (β)
- Operating Characteristic (OC) Function
- Average Sample Number (ASN)
- Efficiency of a Sequential Test
- Wald's SPRT
- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Log-Likelihood Ratio (LLR).

17.9 SELF-ASSESSMENT QUESTIONS:

1. What is the Sequential Probability Ratio Test (SPRT), and how does it differ from a fixed-sample test?
2. Explain the roles of the decision boundaries A and B in the SPRT.
3. What is meant by efficiency in the context of sequential tests?
4. Describe how the test proceeds after each observation in a binomial SPRT.
5. Provide a real-life example where a Poisson-based SPRT may be used.
6. Give an industrial or scientific example where a normal-distribution SPRT is appropriate.
7. Define the log-likelihood ratio (LLR) for sequential testing.
8. What assumptions are needed for the SPRT to be optimal according to Wald?
9. Under what circumstances can the SPRT be significantly more efficient than a fixed-sample test?
10. Why the SPRT is considered asymptotically optimal compared to other sequential tests?

17.10 SUGGESTED READINGS:

1. Goon, A.M., Gupta, M.K., and Dasgupta, B.: Fundamentals of Statistics.
2. Mood, A.M., Graybill, F.A., and Boes, D.C.: Introduction to the Theory of Statistics.
3. Johnson, R.A., and Bhattacharya, G.K.: Statistics-Principles and Methods.
4. Wald, Abraham: Sequential Analysis (a classic text for SPRT).
5. Miller, Irwin and Miller, Marylees: John E. Freund's Mathematical Statistics with Applications.
6. Mathematical Statistics- Parimal Mukopadhyay(1996), New Central Book Agency (P)Ltd., Calcutta.
7. An Introduction to probability and Mathematical Statistics by V.K. Rohatgi.
8. Wetherill, G.B. – *Sequential Methods in Statistics*

Dr. K. Kalyani