# SDS8102: Advanced Mathematical Statistics

## Lecture 5: Foundations of Hypothesis Testing

Saptarshi Roy

Fall 2026

Department of Statistics and Data Science
MBZUAI

# Lecture 5: Hypothesis Testing

Estimation asks:

> What value of the parameter is plausible given the data?

Hypothesis testing asks a different question:

> Is the observed data compatible with a particular claim about the parameter?

Examples include:

- Does a new treatment have an effect?
- Is a coin fair?
- Is a regression coefficient zero?
- Do two populations have the same mean?

Today we develop a mathematical framework for answering such questions.

# Learning Objectives

By the end of this lecture, you should be able to

1. Formulate null and alternative hypotheses.
2. Define a statistical test and rejection region.
3. Distinguish Type I and Type II errors.
4. Define and interpret the power function.
5. Distinguish the size and level of a test.
6. Construct simple tests in the Gaussian model.

# A Running Example: Does a Treatment Work?

Suppose $X_1, \ldots, X_n$ measure the improvement of patients after receiving a treatment.

Assume

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2),$$

where $\sigma^2$ is known.

The treatment has no average effect if

$$\mu = 0.$$

We therefore want to distinguish between

$$\boxed{H_0 : \mu = 0 \qquad \text{versus} \qquad H_1 : \mu > 0.}$$

How should the observed data determine which hypothesis we favor?

# Null and Alternative Hypotheses

Let

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

be a statistical model.

Partition the parameter space into two sets:

$$\Theta_0 \qquad \text{and} \qquad \Theta_1.$$

We wish to test

$$\boxed{H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1.}$$

$H_0$ is called the **null hypothesis** and $H_1$ the **alternative hypothesis**.

# Simple and Composite Hypotheses

A hypothesis is **simple** if it specifies a single distribution.

For example,

$$H_0 : \mu = 0$$

is simple in the model $N(\mu, 1)$.

A hypothesis is **composite** if it contains multiple distributions.

For example,

$$H_1 : \mu > 0$$

is composite because it contains every distribution

$$N(\mu, 1), \qquad \mu > 0.$$

This distinction will become important when we search for **optimal tests**.

# What Is a Statistical Test?

A statistical test is a rule that maps the observed data into a decision.

**Definition: Test Function**

A nonrandomized test is a function

$$\phi : \mathcal{X} \to \{0, 1\},$$

where

$$\phi(X) = \begin{cases} 1, & \text{reject } H_0, \\ 0, & \text{do not reject } H_0. \end{cases}$$

Thus hypothesis testing is itself a statistical decision problem.

# The Rejection Region

Every nonrandomized test can equivalently be described by a **rejection region**

$$R = \{x : \phi(x) = 1\}.$$

The decision rule is then

$$\boxed{X \in R \quad \Longrightarrow \quad \text{reject } H_0.}$$

For our treatment example, a natural rule might be

$$R = \{\overline{X} > c\}.$$

Large positive values of $\overline{X}$ provide evidence against

$$H_0 : \mu = 0$$

in favor of $H_1 : \mu > 0$.

But how should we choose $c$?

# Two Ways a Test Can Be Wrong

There are two possible errors.

| | Do not reject $H_0$ | Reject $H_0$ |
|---|---|---|
| $H_0$ true | Correct | **Type I error** |
| $H_1$ true | **Type II error** | Correct |

A **Type I error** occurs when we reject a true null hypothesis.

A **Type II error** occurs when we fail to reject a false null hypothesis.

These two errors compete with one another.

Making rejection easier reduces Type II error but increases the chance of Type I error.

# The Fundamental Trade-Off

Return to the test

$$\text{reject } H_0 \quad \text{if} \quad \overline{X} > c.$$

If we decrease $c$, rejection becomes easier:

$$\text{smaller } c \Rightarrow \begin{cases} \text{more Type I errors,} \\ \text{fewer Type II errors.} \end{cases}$$

If we increase $c$, rejection becomes harder:

$$\text{larger } c \Rightarrow \begin{cases} \text{fewer Type I errors,} \\ \text{more Type II errors.} \end{cases}$$

We therefore need a principled way to control one error while optimizing the other.

# The Power Function

For a test $\phi$, define its **power function** by

$$\boxed{\beta_\phi(\theta) = \mathbb{P}_\theta(\phi(X) = 1) = \mathbb{E}_\theta[\phi(X)].}$$

Thus $\beta_\phi(\theta)$ is the probability that the test rejects $H_0$ when the true parameter is $\theta$.

Its interpretation depends on where $\theta$ lies:

$$\begin{aligned} \theta \in \Theta_0 \quad &: \quad \text{probability of Type I error,} \\ \theta \in \Theta_1 \quad &: \quad \text{probability of correctly rejecting } H_0. \end{aligned}$$

# Power in the Gaussian Example

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

and reject $H_0 : \mu = 0$ when

$$\overline{X} > c.$$

Since

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

the power function is

$$\begin{aligned}\beta(\mu) &= P_\mu(\overline{X} > c) \\ &= P_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma}\right).\end{aligned}$$

Therefore,

$$\boxed{\beta(\mu) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).}$$

# What Should a Good Power Function Look Like?

For

$$H_0 : \mu \leq 0 \qquad \text{versus} \qquad H_1 : \mu > 0,$$

we want two things.

Under the null,

$$\beta(\mu) \quad \text{should be small.}$$

Under the alternative,

$$\beta(\mu) \quad \text{should be large.}$$

Ideally,

$$\beta(\mu) \approx \begin{cases} 0, & \mu \leq 0, \\ 1, & \mu > 0. \end{cases}$$

Finite samples prevent this ideal separation.

**The goal of test construction is therefore to obtain the best possible power while**

# Size of a Test

For a composite null

$$H_0 : \theta \in \Theta_0,$$

the probability of rejection may differ across values of $\theta$ in $\Theta_0$.

The **size** of a test $\phi$ is

$$\boxed{\text{size}(\phi) = \sup_{\theta \in \Theta_0} \beta_\phi(\theta).}$$

Thus the size is the **largest probability of Type I error** over all distributions satisfying the null hypothesis.

# Level of a Test

A test is said to have **level** $\alpha$ if

$$\boxed{\sup_{\theta \in \Theta_0} P_\theta(\text{reject } H_0) \leq \alpha.}$$

Equivalently,

$$\text{size}(\phi) \leq \alpha.$$

Therefore:

- **size** is a property of a particular test;
- **level** $\alpha$ is a constraint imposed on the test.

A level-0.05 test need not have size exactly 0.05.

It may have size strictly smaller than 0.05.

# Choosing the Critical Value

Return to

$$H_0 : \mu \leq 0 \qquad \text{versus} \qquad H_1 : \mu > 0.$$

Reject when

$$\overline{X} > c.$$

The rejection probability increases with $\mu$, so under $H_0$ the largest rejection probability occurs at the boundary

$$\mu = 0.$$

To construct a level-$\alpha$ test, choose $c$ such that

$$P_{\mu=0}(\overline{X} > c) = \alpha.$$

Since

$$\frac{\sqrt{n}\,\overline{X}}{\sigma} \sim N(0,1) \Longrightarrow \boxed{c = z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.}$$

# The Level-$\alpha$ Gaussian Test

We therefore reject

$$H_0 : \mu \leq 0$$

whenever

$$\boxed{\frac{\sqrt{n}\,\overline{X}}{\sigma} > z_{1-\alpha}.}$$

Its power function is

$$\boxed{\beta(\mu) = 1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{n}\,\mu}{\sigma}\right).}$$

This formula reveals how power depends on

- the significance level $\alpha$,
- the effect size $\mu$,
- the sample size $n$,
- the noise level $\sigma$.

# But Why This Test Statistic?

For the Gaussian problem we chose

$$T(X) = \frac{\sqrt{n}\,\overline{X}}{\sigma}.$$

This choice seems natural because $\overline{X}$ contains information about $\mu$.

But in a general statistical model, there may be many possible test statistics and many possible rejection regions.

Among all level-$\alpha$ tests, which one has the largest power?

This is an **optimization problem over statistical tests**.

The likelihood provides a remarkably powerful answer.

# Simple versus Simple Hypotheses

Consider testing two completely specified distributions:

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1.$$

Let their densities be

$$p_{\theta_0}(x) \qquad \text{and} \qquad p_{\theta_1}(x).$$

For an observed value $X = x$, a natural question is:

Which hypothesis makes the observed data more likely?

This suggests comparing the two likelihoods.

# The Likelihood Ratio

Define the likelihood ratio

$$\Lambda(x) = \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}.$$

Its interpretation is immediate:

$$\Lambda(x) > 1$$

means that the observed data are more likely under $H_1$ than under $H_0$.

More generally, large values of

$$\frac{p_{\theta_1}(X)}{p_{\theta_0}(X)}$$

provide evidence in favor of $H_1$.

This suggests rejection regions of the form

$$\boxed{R = \{x : \Lambda(x) > c\}.}$$

# The Optimization Problem

For a test $\phi$, under the simple null the Type I error is

$$\mathbb{E}_{\theta_0}[\phi(X)].$$

Its power against $\theta_1$ is

$$\mathbb{E}_{\theta_1}[\phi(X)].$$

We therefore want to solve

$$\begin{array}{ll} \text{maximize} & \mathbb{E}_{\theta_1}[\phi(X)] \\ \text{subject to} & \mathbb{E}_{\theta_0}[\phi(X)] \leq \alpha. \end{array}$$

In words:

Among all tests with Type I error at most $\alpha$, find the test with the largest probability of detecting the alternative.

# Most Powerful Tests

**Definition: Most Powerful Test**

A level-$\alpha$ test $\phi^*$ is **most powerful** against $\theta_1$ if

$$\mathbb{E}_{\theta_1}[\phi^*(X)] \geq \mathbb{E}_{\theta_1}[\phi(X)]$$

for every other level-$\alpha$ test $\phi$.

Thus a most powerful test solves an optimization problem over all possible statistical decision rules.

The Neyman–Pearson lemma tells us exactly what this optimal test looks like for simple versus simple hypotheses.

**Theorem (Neyman–Pearson Lemma)**

*Consider*

$$H_0 : \theta = \theta_0 \qquad \textit{versus} \qquad H_1 : \theta = \theta_1.$$

*Suppose a level-$\alpha$ test $\phi^*$ has the form*

$$\phi^*(x) = \begin{cases} 1, & p_{\theta_1}(x) > k\, p_{\theta_0}(x), \\ 0, & p_{\theta_1}(x) < k\, p_{\theta_0}(x), \end{cases}$$

*with possible randomization when equality holds.*

*Then $\phi^*$ is a most powerful level-$\alpha$ test.*

Equivalently, reject $H_0$ for sufficiently large values of

$$\frac{p_{\theta_1}(X)}{p_{\theta_0}(X)}.$$

# Why Should the Likelihood Ratio Be Optimal?

Imagine that we can spend only an $\alpha$ amount of probability under $H_0$ on the rejection region.

Which observations should we put into that region?

We should choose observations for which

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}$$

is largest.

Such observations are

- relatively unlikely under $H_0$,
- relatively likely under $H_1$.

Thus the likelihood-ratio test spends the available Type I error budget where it produces the greatest gain in power.

# Proof of the Neyman–Pearson Lemma

Let $\phi^*$ be the likelihood-ratio test and let $\phi$ be any other level-$\alpha$ test.

By construction,

$$\big(\phi^*(x) - \phi(x)\big)\big(p_{\theta_1}(x) - k p_{\theta_0}(x)\big) \geq 0.$$

Integrating,

$$\int (\phi^* - \phi)(p_{\theta_1} - k p_{\theta_0})\, dx \geq 0.$$

Therefore,

$$\mathbb{E}_{\theta_1}[\phi^*] - \mathbb{E}_{\theta_1}[\phi] \geq k\left(\mathbb{E}_{\theta_0}[\phi^*] - \mathbb{E}_{\theta_0}[\phi]\right).$$

If $\phi^*$ has size $\alpha$ and $\phi$ has level $\alpha$, then

$$\mathbb{E}_{\theta_0}[\phi^*] - \mathbb{E}_{\theta_0}[\phi] \geq 0.$$

Hence

$$\boxed{\mathbb{E}_{\theta_1}[\phi^*] \geq \mathbb{E}_{\theta_1}[\phi].}$$

Thus $\phi^*$ is most powerful. □

# Example: Testing a Gaussian Mean

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2),$$

where $\sigma^2$ is known.

Consider

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu = \mu_1,$$

where

$$\mu_1 > \mu_0.$$

By the Neyman–Pearson lemma, the most powerful test rejects for large values of

$$\frac{L(\mu_1; X)}{L(\mu_0; X)}.$$

What does this likelihood ratio look like?

# The Gaussian Likelihood Ratio

Ignoring constants,

$$L(\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right\}.$$

Hence

$$\log\frac{L(\mu_1)}{L(\mu_0)} = -\frac{1}{2\sigma^2}\left[\sum_i (X_i-\mu_1)^2 - \sum_i (X_i-\mu_0)^2\right].$$

Simplifying,

$$\log\frac{L(\mu_1)}{L(\mu_0)} = \frac{n(\mu_1-\mu_0)}{\sigma^2}\overline{X} - \frac{n(\mu_1^2-\mu_0^2)}{2\sigma^2}.$$

**Since $\mu_1 > \mu_0$, the likelihood ratio is increasing in $\overline{X}$.**

# The Most Powerful Gaussian Test

The Neyman–Pearson rejection region therefore has the form

$$\overline{X} > c.$$

Choose $c$ so that the test has size $\alpha$:

$$P_{\mu_0}(\overline{X} > c) = \alpha.$$

Since

$$\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \sim N(0,1)$$

under $H_0$, we obtain

$$\boxed{\text{Reject } H_0 \quad \text{if} \quad \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} > z_{1-\alpha}.}$$

This is not merely a reasonable test.

It is the **most powerful level-$\alpha$ test** against $\mu = \mu_1$.

# Uniformly most powerful test

For every fixed alternative

$$\mu_1 > \mu_0,$$

the Neyman–Pearson lemma gave the same form of rejection region:

$$\overline{X} > c.$$

Moreover, after imposing level $\alpha$, the critical value $c$ depends on $\mu_0$, $\sigma$, $n$, and $\alpha$—but **not on** $\mu_1$.

Therefore the same test is most powerful against every alternative

$$\mu_1 > \mu_0.$$

This suggests a stronger notion of optimality:

Can one test be most powerful simultaneously against every alternative?

Such a test is called **uniformly most powerful**.

# Uniformly Most Powerful Tests

Consider

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1.$$

**Definition: Uniformly Most Powerful Test**

A level-$\alpha$ test $\phi^*$ is **uniformly most powerful (UMP)** if, for every level-$\alpha$ test $\phi$,

$$\boxed{\beta_{\phi^*}(\theta) \geq \beta_{\phi}(\theta) \qquad \text{for every } \theta \in \Theta_1.}$$

Thus $\phi^*$ is simultaneously most powerful against **every parameter value in the alternative**.

# Most Powerful versus Uniformly Most Powerful

A most powerful test solves a simple-versus-simple problem:

$$\theta_0 \quad \text{versus} \quad \theta_1.$$

A UMP test must solve many such problems simultaneously:

$$\theta_0 \quad \text{versus} \quad \theta \qquad \text{for every } \theta \in \Theta_1.$$

Therefore,

| UMP is a much stronger requirement than most powerful. |
|---|

Different alternatives may favor different rejection regions.

For this reason, UMP tests do not exist in every testing problem.

# Why Did a UMP Test Exist for the Gaussian Mean?

Recall

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

and

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu > \mu_0.$$

For every $\mu_1 > \mu_0$, we found

$$\frac{L(\mu_1; X)}{L(\mu_0; X)}$$

to be an increasing function of $\overline{X}$.

Thus every simple alternative $\mu_1 > \mu_0$ favors the **same ordering of the sample space**:

$$\boxed{\text{larger } \overline{X} \quad \Longrightarrow \quad \text{stronger evidence for larger } \mu.}$$

This common ordering is the key.

# Monotone Likelihood Ratio

**Definition: Monotone Likelihood Ratio**

A family $\{P_\theta : \theta \in \Theta\}$ has a **monotone likelihood ratio** in a statistic $T(X)$ if, whenever

$$\theta_1 > \theta_0,$$

the likelihood ratio

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}$$

is a nondecreasing function of $T(x)$.

In words:

Larger $T(X)$ systematically favors larger values of $\theta$.

## Why MLR Is Powerful

Suppose the family has monotone likelihood ratio in $T(X)$.

For any

$$\theta_1 > \theta_0,$$

the Neyman–Pearson lemma says to reject $H_0 : \theta = \theta_0$ for large values of

$$\frac{p_{\theta_1}(X)}{p_{\theta_0}(X)}.$$

But MLR implies that this is equivalent to rejecting for large values of

$$T(X).$$

Crucially, the ordering by $T(X)$ is the same for every $\theta_1 > \theta_0$.

This is what makes uniform optimality possible.

# UMP Tests from Monotone Likelihood Ratio

**Theorem (Karlin–Rubin)**

*Suppose* $\{P_\theta : \theta \in \Theta\}$ *has monotone likelihood ratio in* $T(X)$.

*For testing*

$$H_0 : \theta \leq \theta_0 \qquad \textit{versus} \qquad H_1 : \theta > \theta_0,$$

*a test of the form*

$$\boxed{\phi(X) = \mathbf{1}\{T(X) > c\}}$$

*with* $c$ *chosen to have level* $\alpha$ *is UMP among level-*$\alpha$ *tests.*

For a decreasing likelihood ratio, the direction of the rejection region is reversed.

# Why the Boundary of the Null Matters

Suppose the rejection region is

$$T(X) > c.$$

Under monotone likelihood ratio, the rejection probability

$$P_\theta(T(X) > c)$$

is nondecreasing in $\theta$.

Therefore, under

$$H_0 : \theta \leq \theta_0,$$

the largest Type I error occurs at

$$\boxed{\theta = \theta_0.}$$

Hence $c$ can be chosen by solving

$$P_{\theta_0}(T(X) > c) = \alpha.$$

# Example: Gaussian Mean Has MLR

For

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2),$$

consider $\mu_1 > \mu_0$.

We previously obtained

$$\log \frac{L(\mu_1)}{L(\mu_0)} = \frac{n(\mu_1 - \mu_0)}{\sigma^2} \overline{X} - \frac{n(\mu_1^2 - \mu_0^2)}{2\sigma^2}.$$

Since

$$\frac{n(\mu_1 - \mu_0)}{\sigma^2} > 0,$$

the likelihood ratio is increasing in

$$T(X) = \overline{X}.$$

Therefore the Normal location family has monotone likelihood ratio in $\overline{X}$.

# UMP Test for the Gaussian Mean

Consider

$$H_0 : \mu \leq \mu_0 \qquad \text{versus} \qquad H_1 : \mu > \mu_0.$$

By the monotone likelihood ratio property, the UMP level-$\alpha$ test rejects for large $\overline{X}$.

Choosing the critical value at $\mu = \mu_0$ gives

$$\boxed{\text{Reject } H_0 \quad \text{if} \quad \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} > z_{1-\alpha}.}$$

Thus the familiar one-sided $z$-test is not merely intuitive:

$$\boxed{\text{It is UMP among all level-}\alpha\text{ tests.}}$$

# Example: Bernoulli Model

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p).$$

Let

$$T = \sum_{i=1}^{n} X_i.$$

For $p_1 > p_0$,

$$\frac{L(p_1)}{L(p_0)} = \left(\frac{p_1}{p_0}\right)^T \left(\frac{1-p_1}{1-p_0}\right)^{n-T}.$$

Rearranging,

$$\frac{L(p_1)}{L(p_0)} = \left(\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right)^T \left(\frac{1-p_1}{1-p_0}\right)^n.$$

Since $p_1 > p_0$, this is increasing in $T$.

# UMP Test for a Binomial Probability

Therefore the Bernoulli family has monotone likelihood ratio in

$$T = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p_0)$$

For testing

$$H_0 : p \leq p_0 \qquad \text{versus} \qquad H_1 : p > p_0,$$

the UMP level-$\alpha$ test rejects for large values of $T$:

$$\boxed{T \geq c.}$$

The critical value is chosen so that

$$P_{p_0}(T \geq c) \leq \alpha \quad (c \text{ can be chosen using Binomial quantiles.})$$

# A Subtlety: Discrete Test Statistics

Suppose

$$T \sim \text{Binomial}(n, p_0).$$

Because $T$ takes only integer values, there may be no critical value $c$ satisfying

$$P_{p_0}(T \geq c) = \alpha$$

exactly.

We can choose $c$ so that

$$P_{p_0}(T \geq c) \leq \alpha,$$

giving a valid level-$\alpha$ test.

But its size may be strictly smaller than $\alpha$.

This is one reason randomized tests appear naturally in the general theory.

## What About Two-Sided Alternatives?

So far the alternatives have had a preferred direction:

$$H_1 : \theta > \theta_0.$$

Monotone likelihood ratio tells us that large $T(X)$ provides evidence in that direction.

Now consider

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

For $\theta > \theta_0$, we want to reject for **large** $T(X)$.

But for $\theta < \theta_0$, we typically want to reject for **small** $T(X)$.

Thus the most powerful rejection region depends on which side of the null the alternative lies.

A UMP test may therefore fail to exist for two-sided alternatives.

# The Difficulty with Two-Sided Alternatives

Consider

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

Suppose the model has monotone likelihood ratio in $T(X)$.

For alternatives

$$\theta > \theta_0,$$

the Neyman–Pearson principle favors rejection for **large** values of $T(X)$.

But for alternatives

$$\theta < \theta_0,$$

it favors rejection for **small** values of $T(X)$.

Thus the two directions compete for the available Type I error.

# Example: Two-Sided Gaussian Mean

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2),$$

where $\sigma^2$ is known.

Consider

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0.$$

For any fixed $\mu_1 > \mu_0$, the most powerful level-$\alpha$ test rejects for large values of $\overline{X}$.

For any fixed $\mu_1 < \mu_0$, the most powerful level-$\alpha$ test rejects for small values of $\overline{X}$.

The most powerful rejection regions point in opposite directions.

# Why a UMP Test Does Not Exist

Suppose a UMP level-$\alpha$ test existed for

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0.$$

It would have to be most powerful against every $\mu_1 > \mu_0$.

By the Neyman–Pearson lemma, this requires spending the Type I error in the **upper tail**.

But it must also be most powerful against every $\mu_1 < \mu_0$, which requires spending the Type I error in the **lower tail**.

These requirements cannot generally be satisfied simultaneously.

There is no UMP level-$\alpha$ test for the usual two-sided Gaussian mean problem.

# A Weaker Requirement: Unbiased Tests

Recall that the power function is

$$\beta_\phi(\theta) = P_\theta(\text{reject } H_0).$$

A natural requirement is that the test should be more likely to reject under the alternative than under the null.

**Definition: Unbiased Test**

A level-$\alpha$ test is **unbiased** if

$$\boxed{\beta_\phi(\theta) \geq \alpha \qquad \text{for every } \theta \in \Theta_1.}$$

Thus moving away from the null should not make rejection *less likely* than its nominal level.

## Why Unbiasedness Helps

For the two-sided problem

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0,$$

an unbiased level-$\alpha$ test satisfies

$$\beta(\mu_0) = \alpha$$

and

$$\beta(\mu) \geq \alpha \qquad \text{for every } \mu \neq \mu_0.$$

Thus the power function has its minimum at the null:

$$\boxed{\mu_0 \text{ is a minimizer of } \beta(\mu).}$$

For a smooth power function, this suggests the condition

$$\beta'(\mu_0) = 0.$$

# Uniformly Most Powerful Unbiased Tests

Once we restrict attention to unbiased tests, we can again ask for optimality.

**Definition: UMPU Test**

A level-$\alpha$ test $\phi^*$ is **uniformly most powerful unbiased (UMPU)** if it is unbiased and

$$\boxed{\beta_{\phi^*}(\theta) \geq \beta_{\phi}(\theta) \qquad \text{for every } \theta \in \Theta_1}$$

for every other unbiased level-$\alpha$ test $\phi$.

The restriction to unbiased tests makes optimal two-sided procedures possible in many important models.

# The Two-Sided Gaussian Test

For

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0,$$

define

$$Z = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}.$$

Under $H_0$,

$$Z \sim N(0, 1).$$

The familiar two-sided level-$\alpha$ test rejects when

$$\boxed{|Z| > z_{1-\alpha/2}.}$$

Equivalently,

$$Z < -z_{1-\alpha/2} \qquad \text{or} \qquad Z > z_{1-\alpha/2}.$$

The Type I error $\alpha$ is split equally between the two tails.

# Power of the Two-Sided Gaussian Test

Under a general value $\mu$,

$$Z \sim N\left(\frac{\sqrt{n}(\mu - \mu_0)}{\sigma}, 1\right).$$

Let

$$\delta = \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}.$$

The power is

$$\begin{aligned}\beta(\mu) &= P_\mu(|Z| > z_{1-\alpha/2}) \\ &= P_\mu(Z < -z_{1-\alpha/2}) + P_\mu(Z > z_{1-\alpha/2}).\end{aligned}$$

Therefore,

$$\boxed{\beta(\mu) = \Phi(-z_{1-\alpha/2} - \delta) + 1 - \Phi(z_{1-\alpha/2} - \delta).}$$

# The Shape of the Power Function

At the null,

$$\beta(\mu_0) = \alpha.$$

As $\mu$ moves away from $\mu_0$ in either direction,

$$\beta(\mu) \longrightarrow 1.$$

Moreover, the power function is symmetric around $\mu_0$:

$$\boxed{\beta(\mu_0 + h) = \beta(\mu_0 - h).}$$

Thus

$$\beta'(\mu_0) = 0,$$

and the test is unbiased.

In the Normal model with known variance, this familiar two-sided test is in fact **UMPU**.

# A Hierarchy of Optimal Tests

We have now encountered three notions of optimality.

| Setting | Optimality | Main tool |
|---|---|---|
| Simple vs. simple | Most powerful | Neyman–Pearson |
| One-sided composite | UMP | Monotone likelihood ratio |
| Two-sided composite | Often no UMP | UMP unbiased |

The key distinction is:

> One-sided alternatives have a direction; two-sided alternatives compete across directions.

This is why two-sided optimal testing requires additional structure or restrictions.