MATH 5010 –Foundations of Statistical Theory and Probability

❖ Hypothesis Testing 2

-- Evaluating Tests

Instructor: He Wang
Department of Mathematics
Northeastern University

Evaluating Hypothesis Tests

How do we decide if one test is *better* than another?

In Hypothesis, we may make a mistake. Hypothesis tests are evaluated and compared through their probabilities of making mistakes.

Determine which tests have the smallest possible error probabilities.

- Power Function
- Size (Significance Level)
- Unbiased Tests
- Uniformly Most Powerful (UMP) Test
- Risk Function (Expected Loss)

Two Types of Errors in Hypothesis testing

Hypothesis Test $\boldsymbol{H_0}: \theta \in \Theta_0$ v.s. $\boldsymbol{H}_1: \theta \in \Theta_0^c$

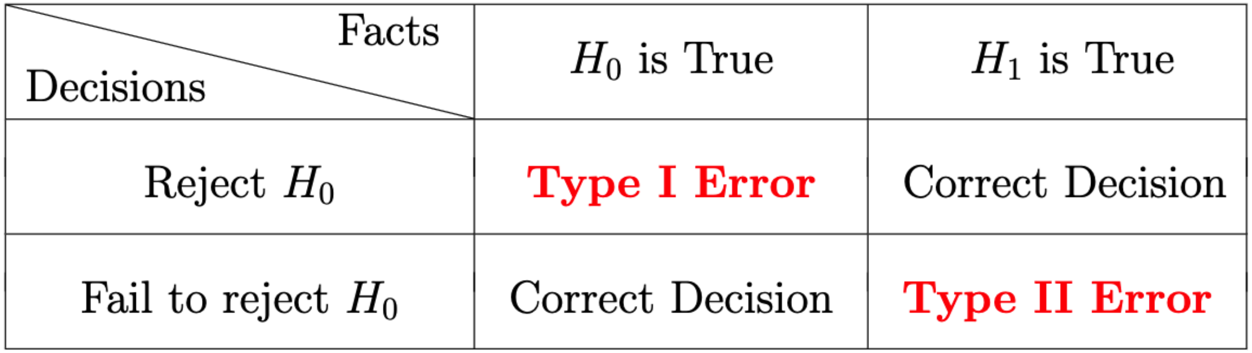
Two types of errors:

Type I Error:

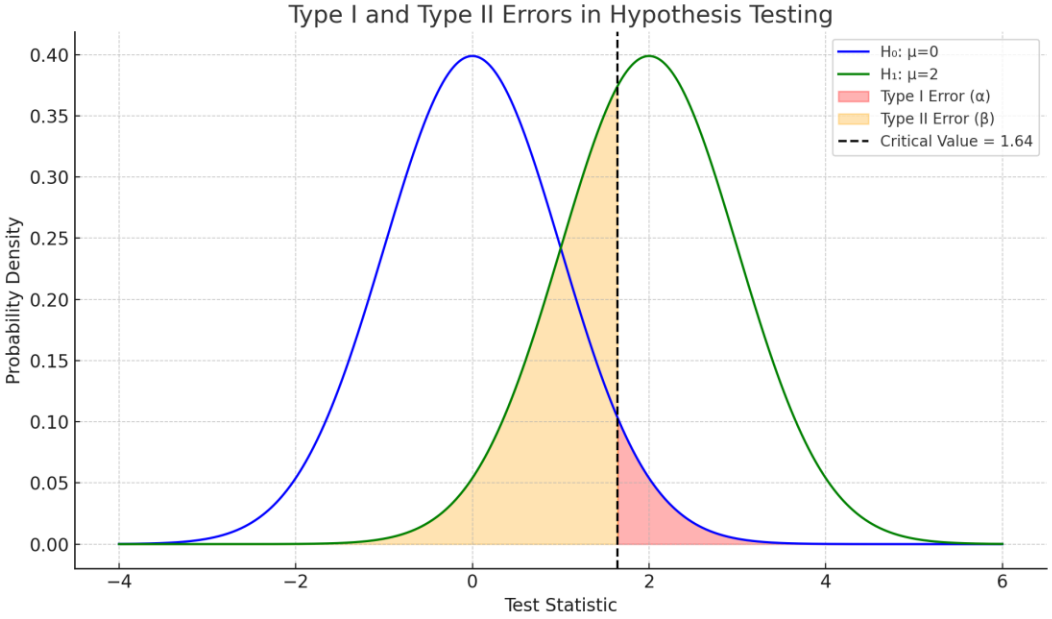
If $\theta \in \Theta_0$, but the test decides to reject H_0.

Type II Error:

If $\theta \in \Theta_0^c$, but the test decides to accept H_0



Facts Decisions	H_0 is True	H_1 is True
Reject H_0	**Type I Error**	Correct Decision
Fail to reject H_0	Correct Decision	**Type II Error**


Type I and Type II Errors in Hypothesis Testing
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Probability Density
−4
−2
0
2
4
6
Test Statistic
H₀: μ=0
H₁: μ=2
Type I Error (α)
Type II Error (β)
Critical Value = 1.64

Error Probabilities (Power function)

Suppose R is the rejection region for a test $\theta \in \Theta_0 \ v.s. \theta \in \Theta_0^c$.

For $\theta \in \Theta_0$, the probability of type I error is $P_\theta(\vec{X} \in R)$

For $\theta \in \Theta_0^c$, the probability of type II error is $P_\theta(\vec{X} \in R^c) = 1 - P_\theta(\vec{X} \in R\ )$

Definition: The **power function** of a hypothesis test with rejection region R is the function of θ defined by

$$\beta(\theta) \coloneqq P_\theta(\vec{X} \in R) = \begin{cases} P(\text{Type I Error}) & \text{if } \theta \in \Theta_0 \\ 1 - P(\text{Type II Error}) & \text{if } \theta \in \Theta_0^c \end{cases}$$

Ideal power function (no error):

$$\beta(\theta) = \begin{cases} 0 \text{ if } \theta \in \Theta_0 \\ 1 \text{ if } \theta \in \Theta_0^c \end{cases}$$

Example: Binomial Power function:

$X \sim Binomial(n = 5, \theta)$

Hypothesis $H_0 \colon \theta \leq \frac{1}{2}$ v.s. $H_1 \colon \theta > \frac{1}{2}$

Test 1 – Reject H_0 if $X = 5$.

Power function:

$$\beta_1(\theta) = P_\theta\left(\vec{X} \in R\right) = P(\mathrm{X} = 5) = \theta^5$$

Type I error: For $\theta \leq \frac{1}{2}, \beta_1(\theta) \leq \left(\frac{1}{2}\right)^5 = 0.031$

Type II error: For $\theta > \frac{1}{2}, 1 - \beta_1(\theta) > 1 - \left(\frac{1}{2}\right)^5$ (large for most $\theta > \frac{1}{2}$)

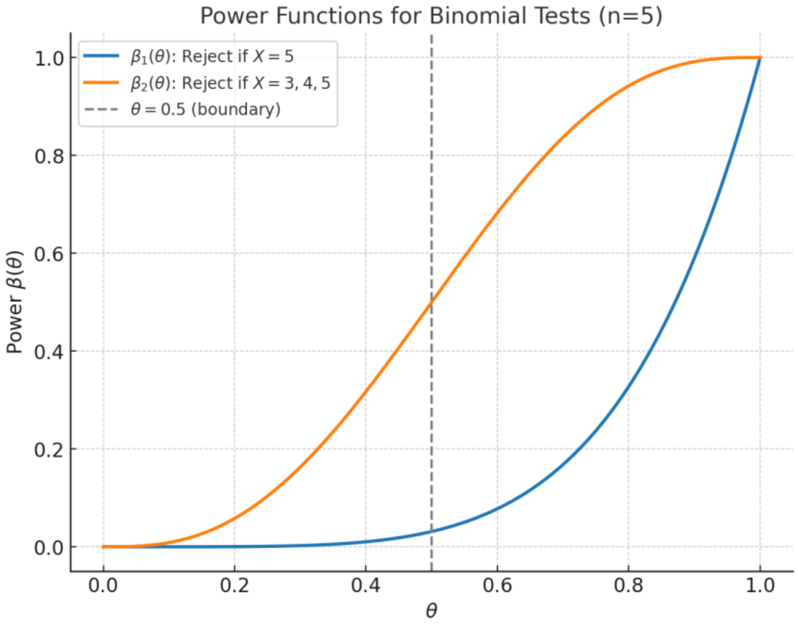
Test 2 – Reject H_0 if $X = 3{,}4, or$ 5

Power function:

$$\beta_2(\theta) = P_\theta(X = 3, 4, 5) = \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta) + \binom{5}{5}\theta^5.$$

Type I Error: Larger than in Test 1

Type II Error: Smaller compared to Test 1



Example: Normal power function

Sample: $X_1, \dots, X_n \sim Normal(\theta, \sigma^2)$ with σ^2 known.

Hypotheses $H_0: \theta \leq \theta_0$ v.s. $H_1: \theta > \theta_0$

Test: Reject H_0 if $\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c$ for a constant $c > 0$

Power function

$$\beta(\theta) = P_\theta\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\right)$$

$$= P_\theta\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$

$$\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} = Z \sim N(0,1)$$

$$= P_\theta\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$

- As $\theta \to -\infty, \;\; \beta(\theta) \to 0$
- As $\theta \to +\infty, \;\; \beta(\theta) \to 1$
- At $\theta = \theta_0$,

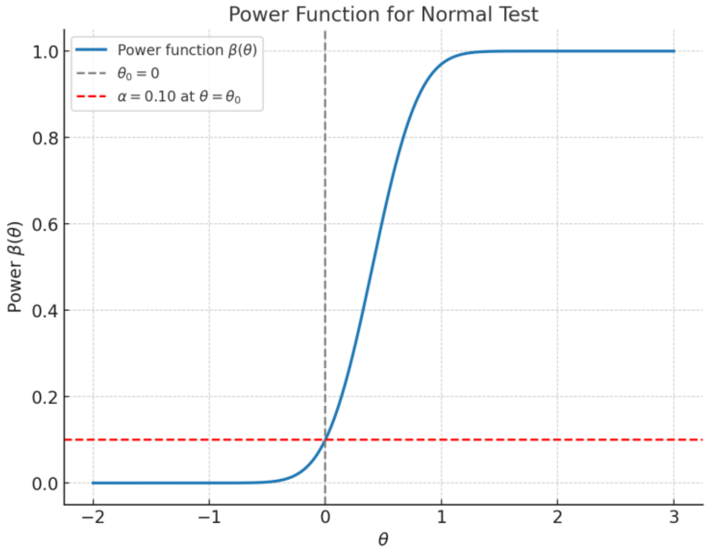
$$\beta(\theta_0) = P(Z > c) = \alpha$$

For example $c = 1.28, \; P(Z > 1.28) \approx 0.10 = 10\%$ significance level

$\theta_0 = 0$

$n = 10$

$\sigma = 1$



For a fixed sample size, it is a trade-off between two types of errors. Usually impossible to make both types of error probabilities very small.

Example: Normal

Sample: $X_1, \dots, X_n \sim Normal(\theta, \sigma^2)$ with σ^2 known.

Hypotheses $H_0: \theta \leq \theta_0$ v.s. $H_1: \theta > \theta_0$

Test: Reject H_0 if $\dfrac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c$ for a constant $c > 0$

Choose $\boldsymbol{c\ and\ n}$ to achieve

$$P(type\ I\ error) \leq 0.1$$

$P(type\ II\ error) \leq 0.2$ if $\theta \geq \theta_0 + \sigma$

Power function:

$$\beta(\theta) = P_\theta\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$

We need to solve

$$\beta(\theta_0) = 0.1$$

$$\beta(\theta_0 + \sigma) = 0.8$$

We know that $c = 1.28$, $P(Z > 1.28) \approx 0.10 = 10\%$ significance level

$$\beta(\theta_0 + \sigma) = P_\theta\left(Z > c - \frac{1}{1/\sqrt{n}}\right)$$

$$= P_\theta\ (Z > 1.28 - \sqrt{n}) = 0.8$$

Solve $n \approx 4.49$. So $n = 5$.

Size α test

Definition: For $\alpha \in [0,1]$, a test with power function $\beta(\theta)$ is called a **size $\boldsymbol{\alpha}$ test**, if

$$\sup_{\theta \in \Theta_0} P_\theta(\text{reject } H_0) = \alpha$$

That is

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$$

A test is called **level $\boldsymbol{\alpha}$ test**, if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

Size of Likelihood Ratio Test (LRT)

For a likelihood-ratio test (LRT) with rejection region $\{\lambda(x) \le c\}$, the **size** is

$$\sup_{\theta \in \Theta_0} P_\theta(\text{reject } H_0) = \alpha$$

So we choose the cutoff c so that the **worst-case** (largest) rejection probability under H_0 equals α.

For example, $H_0: \theta = \theta_0$ and $\frac{(\bar{X} - \theta_0)}{\sigma/\sqrt{n}} \sim Normal(0,1)$

$$Reject\ H_0\ if\ \left|\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}\right| \ge z_{\alpha/2}$$

It corresponds to $c = \exp(-z_{\alpha/2}^2/2)$ for LRT statistic.

Size-α LRT

Standard "critical values" used to state rejection rules

$$z_\alpha\text{: } P(Z > z_\alpha) = \alpha \quad for\ Z \sim N(0,1)$$

$$t_{n-1,\alpha/2}\text{: } P(T_{n-1} > t_{n-1,\alpha/2}) = \alpha/2$$

$$\chi^2_{p,1-\alpha}\text{: } P\left(\chi^2_p > \chi^2_{p,1-\alpha}\right) = \alpha$$

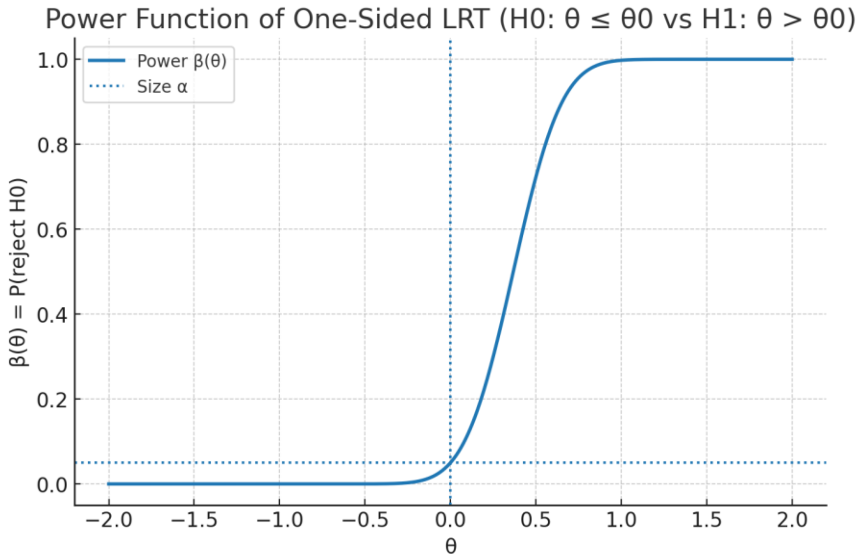
Unbiased Test

A test with power function $\beta(\theta)$ is **unbiased** if

$$\beta(\theta') \geq \beta(\theta'')$$

for every $\theta' \in \Theta_0^c$ and $\theta'' \in \Theta_0$

That is, the probability of rejecting under the alternative must be at least as large as under the null.



Most Powerful Tests:

A *level α test* guarantees that the probability of a Type I Error is at most α for all $\theta \in \Theta_0$.

Next, we also want to control the probability of Type II Error small.

Definition: Let $\mathcal{C}$ be a class of tests for testing

$$H_0: \theta \in \Theta_0 \text{ v.s. } H_1: \theta \in \Theta_0^c$$

A test in $\mathcal{C}$, with power function $\beta(\theta)$, is called a **uniformly most powerful (UMP) test** in $\mathcal{C}$ if

$$\beta(\theta) \geq \beta'(\theta), \text{ for any } \theta \in \Theta_0^c$$

for every competing power function $\beta'(\theta)$ corresponding to another test in $\mathcal{C}$.

Neyman–Pearson Lemma

Consider testing $H_0\colon \theta = \theta_0$ v.s. $H_1\colon \theta = \theta_1$ where the pdf/pmf of X under θ_i is $f(x|\theta_i)$.

Define a **rejection region** R such that:

$$x \in R \text{ if } f(\,x \mid \theta_1\,) > k\,f(x \mid \theta_0),$$

$$x \in R^c \text{ if } f(\,x \mid \theta_1\,) < k\,f(x \mid \theta_0),$$

for some constant $k \geq 0$, and such that

$$P_{\theta_0}(X \in R) = \alpha$$

Then:

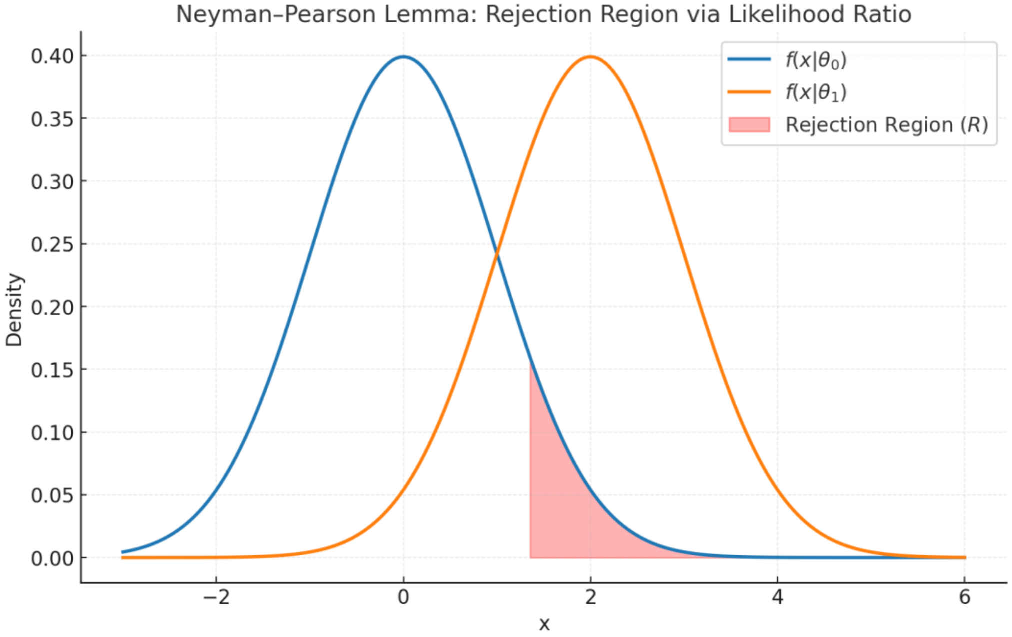
- Any test that satisfies above conditions is a UMP level α test

- If such a test exists with $k > 0$, then every UMP level α test must satisfy the inequality, except possibly on a set A where both hypotheses assign probability zero:

$$P_{\theta_0}(X \in A) = P_{\theta_1}(X \in A) = 0$$

The **most powerful test** for testing $H_0: \theta = \theta_0$ vs $H_1: \theta = \theta_1$ is based on the **likelihood ratio**:

$$\frac{f(x \mid \theta_1)}{f(x \mid \theta_0)} = \frac{L(\theta_1 \mid x)}{L(\theta_0 \mid x)} \geq k$$



The red shaded region represents the rejection region RRR, where the likelihood ratio

$$\frac{f(x \mid \theta_1)}{f(x \mid \theta_0)} > k$$

In this region, evidence in favor of H_1 is strong enough that we reject H_0.

Example: UMP Binomial Test

$$X \sim Binomial(2, \theta)$$

Test $H_0 : \theta = \theta_0 = \frac{1}{2}$ vs $H_1 : \theta = \theta_1 = \frac{3}{4}$

Compute likelihood ratios for $X \in \{0,1,2\}$

$$\frac{f(\,0 \mid \theta = \theta_1\,)}{f(\,0 \mid \theta = \theta_0\,)} = \frac{1}{4} \quad < \quad \frac{f(\,1 \mid \theta = \theta_1\,)}{f(\,1 \mid \theta = \theta_0\,)} = \frac{3}{4} \quad < \quad \frac{f(\,2 \mid \theta = \theta_1\,)}{f(\,2 \mid \theta = \theta_0\,)} = \frac{9}{4}$$

Choose threshold k and define the rejection region by Neyman–Pearson Lemma

- If $\frac{3}{4} < k < \frac{9}{4}$, Reject H_0 when $X = 2$.

This gives a level-α test with $\alpha = P\left(X = 2 \middle| \theta = \frac{1}{2}\right) = \frac{1}{4}$

- If $\frac{1}{4} < k < \frac{3}{4}$, Reject H_0 when $X = 1,2$.

  This gives a level-α test with $\alpha = P\left(X = 1 \; or \; 2 \middle| \theta = \frac{1}{2}\right) = \frac{3}{4}$

- If $\mathrm{k} < \frac{1}{4}$, Reject H_0 when $X = 0,1,2$.

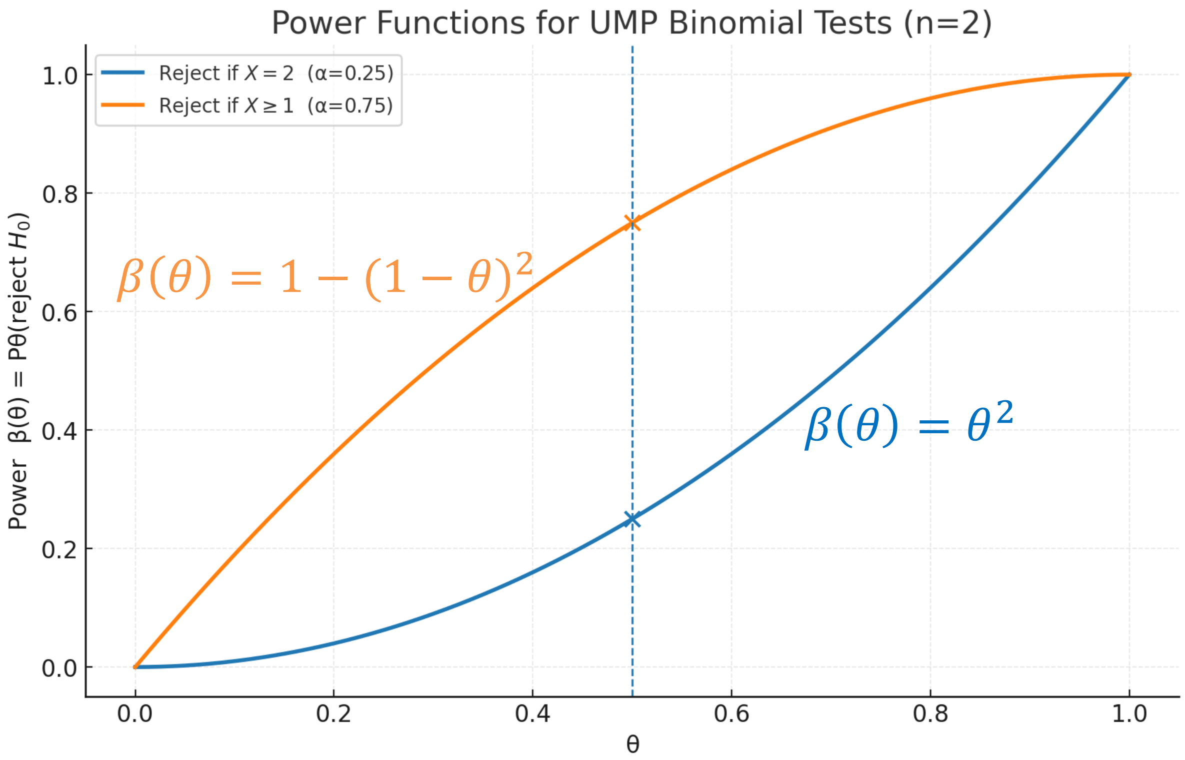
  This gives a level-α test with $\alpha = P\left(X = 0,1,2 \middle| \theta = \frac{1}{2}\right) = 1$

- If $\mathrm{k} > \frac{9}{4}$, Nevel Reject H_0

  This gives a level-α test with $\alpha = 0$

For the boundary, $k = \frac{3}{4}$, we must reject H_0 for $X = 2$ and accept H_0 for $X = 0$.
But the decision for $X = 1$ is ambiguous.

We will get a level-$\alpha = \frac{3}{4}$ test, if reject H_0 when $X = 1$.


Power Functions for UMP Binomial Tests (n=2)
Reject if $X = 2$ (α=0.25)
Reject if $X \geq 1$ (α=0.75)
$\beta(\theta) = 1 - (1 - \theta)^2$
$\beta(\theta) = \theta^2$
Power β(θ) = Pθ(reject H_0)
θ
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

UMP Normal Test

Suppose $X_1, \ldots, X_n \sim Normal(\theta, \sigma^2)$ with known σ^2

The sample mean $\bar{X}$ is a sufficient statistic for θ.

Test $H_0: \theta = \theta_0$ vs $H_1: \theta = \theta_1$, where $\theta_0 > \theta_1$

The Neyman–Pearson test rejects H_0 when

$$g(\bar{x}|\theta_1) > kg(\bar{x}|\theta_0)$$

Equivalently,

$$\bar{x} < \frac{(2\sigma^2 \log k)/n - \theta_0^2 + \theta_1^2}{2(\theta_1 - \theta_0)}$$

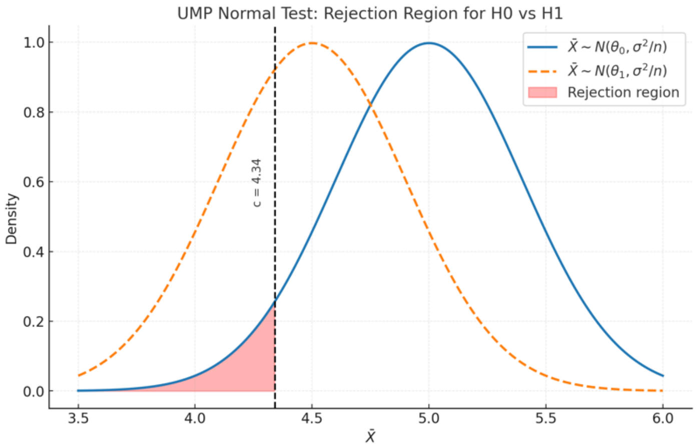
The right-hand side is a increasing function of k, So, the test is

Reject H_0 if $\bar{X} < c$, for some cutoff c

To make this a **UMP level−$\boldsymbol{\alpha}$ test**, we choose c so that

$$\alpha = P_{\theta_0}(\bar{X} < c)$$

Under $H_0, \bar{X} \sim Normal\left(\theta_0, \frac{\sigma^2}{n}\right)$, we set $c = \theta_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$.



➢ $\boldsymbol{p-}$ **Values**

A **p-value**, $p(X)$, is a test statistic that satisfies

$$1 \leq p(x) \leq 1$$

for every sample point x.

Small values of $p(X)$ provide evidence in favor of the alternative hypothesis $H1$

A p-value is considered **valid** if, for every $\theta \in \Theta_0$, and for all $0 \leq \alpha \leq 1$,

$$P_\theta(p(X) \leq \alpha) \leq \alpha$$

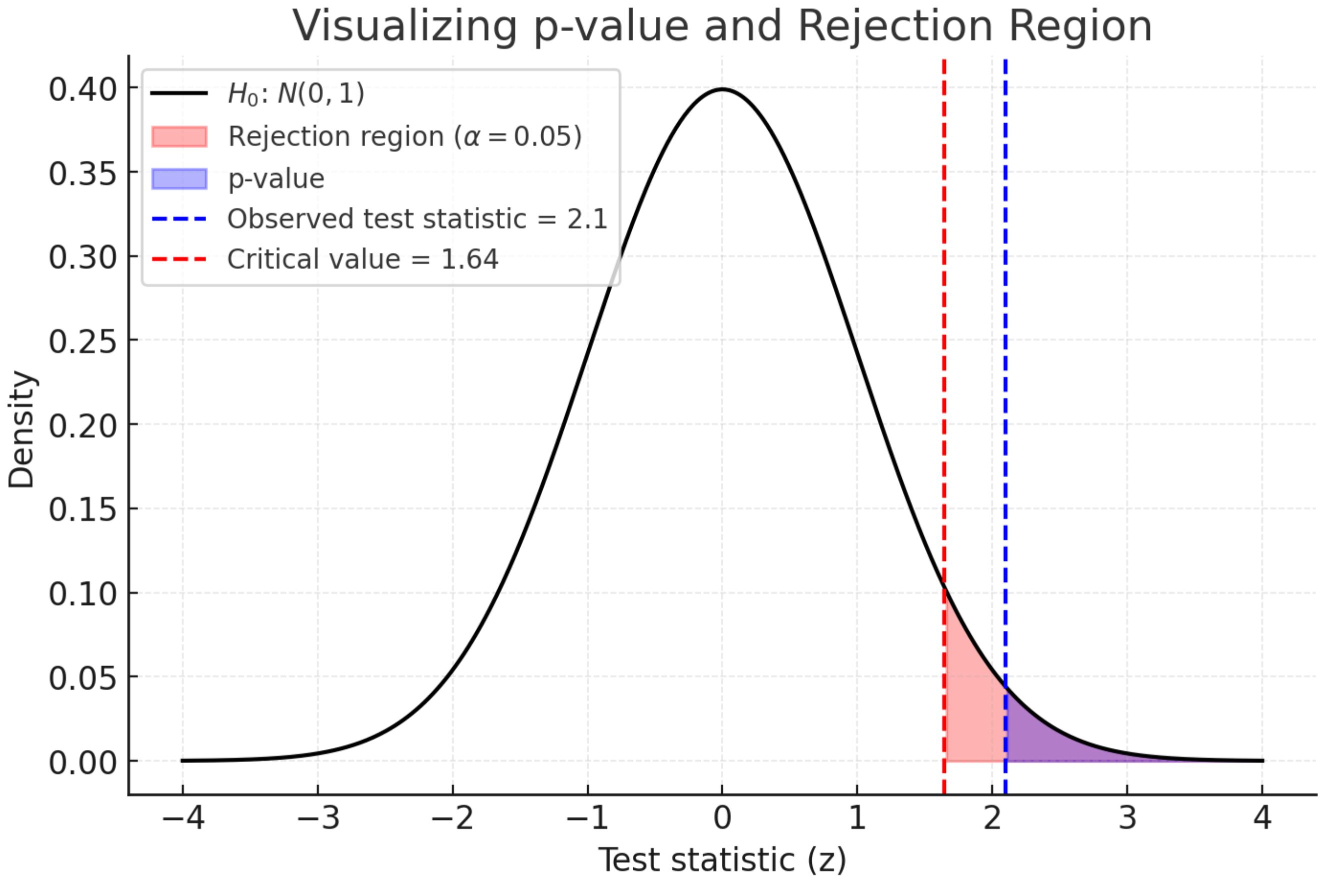
Remark: The p-value measures how compatible the observed data are with the null hypothesis H_0.

The smaller the p-value, the stronger the evidence against the null hypothesis

If $p \leq \alpha$, we reject H_0.

The p-value can be understood as the smallest significance level α at which the observed test statistic leads to rejection of the null hypothesis H_0

$$p(x) = \inf\{\alpha\colon \text{reject } H_0 \; at \; level \; \alpha\}.$$



Theorem:

Let $W(X)$ be a test statistic such that **large values of $\boldsymbol{W}$** provide evidence in favor of the alternative hypothesis H_1.

For each sample point x, the following defines a valid p-value:

$$p(x) \coloneqq \sup_{\theta \in \Theta_0} P_\theta(W(X) \geq W(x))$$

Remark: A p-value is the maximum probability, under the null hypothesis, of observing a test statistic at least as extreme as the one we actually observed.

The p-value measures: *How surprising is my observed signal, if the null hypothesis were true?*

Example: One-sided Normal p-value

Let $X_1, \dots, X_n \sim Normal(\mu, \sigma^2)$ with σ known.

Test $H_0 \colon \mu \leq \mu_0$ vs $H_1 \colon \mu > \mu_0$

Test statistic:

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

Under H_0 with $\mu = \mu_0$, $Z \sim N(0,1)$.

For any $\mu \leq \mu_0$,

$$P_\mu(W \geq w) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq w + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \leq P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq w\right)$$

Thus the supremum over $\Theta_0 = \{\mu \colon \mu \leq \mu_0\}$ is attained at the boundary $\mu = \mu_0$

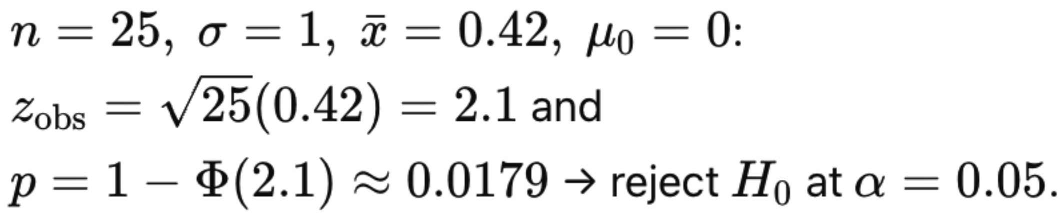
Therefore the valid p-value is

$$p(x) = P_{\mu_0}(Z \geq z_{obs}) = 1 - \Phi(z_{obs})$$

Φ: cdf of Z

where, $z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

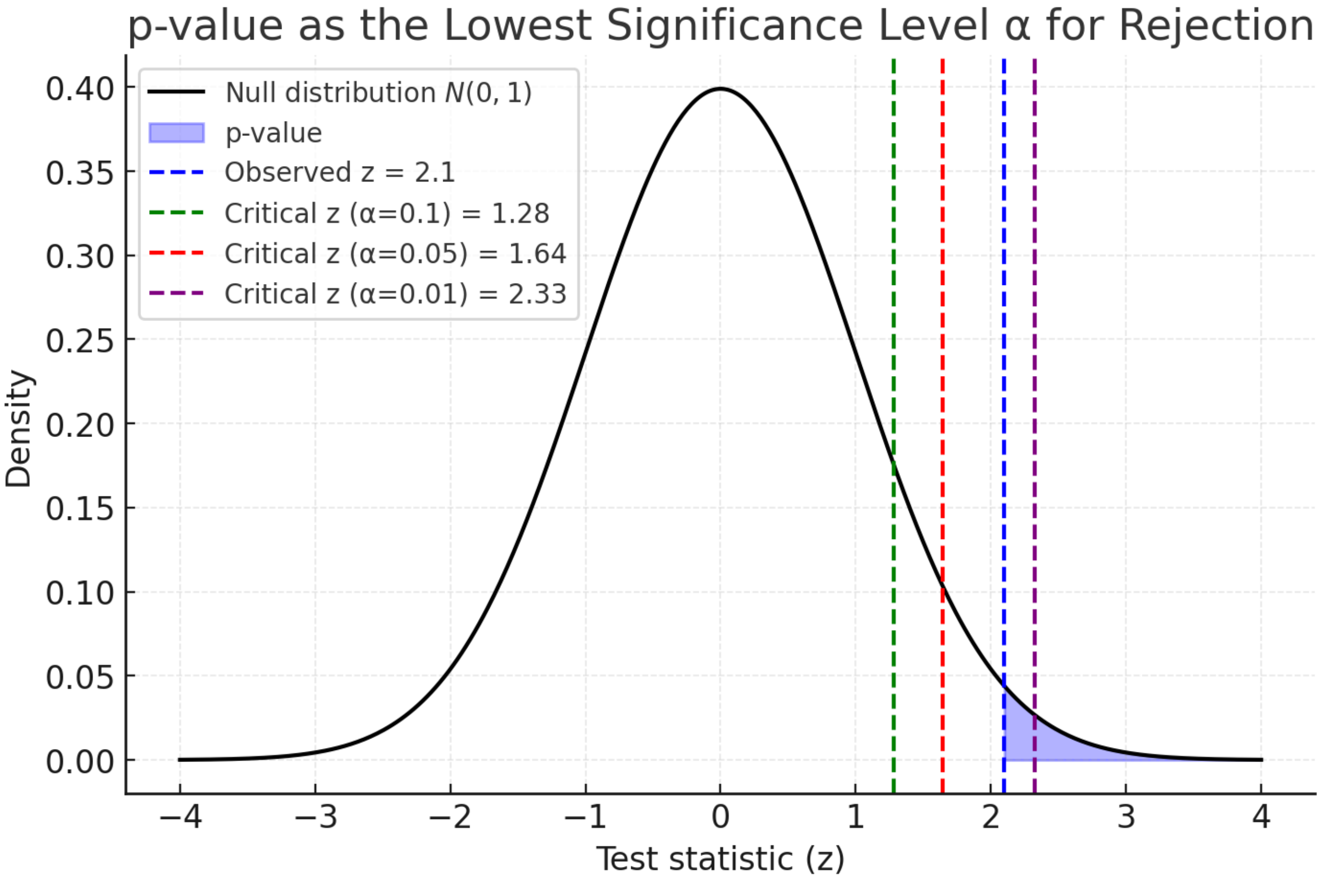
For example,



$n = 25,\ \sigma = 1,\ \bar{x} = 0.42,\ \mu_0 = 0$:

$z_{\text{obs}} = \sqrt{25}(0.42) = 2.1$ and

$p = 1 - \Phi(2.1) \approx 0.0179$ → reject H_0 at $\alpha = 0.05$.


p-value as the Lowest Significance Level α for Rejection
Null distribution $N(0, 1)$
p-value
Observed z = 2.1
Critical z (α=0.1) = 1.28
Critical z (α=0.05) = 1.64
Critical z (α=0.01) = 2.33
Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
Test statistic (z)

- **Loss and Risk functions-** decision theoretic method

In hypothesis testing, there are only two possible **actions**:

A **decision rule** $\delta(x)$ determines whether we accept or reject H_0, based on **data** x.
So, the **decision space** is $\{a_0, a_1\}$

The **loss function** $L(\theta, a)$ measures the "penalty" for taking action a when the true state of nature is θ.

Example: 0–1 loss

$$L(\theta, a) = \begin{cases} 0 & \text{If } \theta \in \Theta_0 \text{ and Accept } H_0\text{, or if } \theta \in \Theta_0^c \text{ and Reject } H_0 \\ 1 & \text{If } \theta \in \Theta_0 \text{ and Reject } H_0\text{, or if } \theta \in \Theta_0^c \text{ and Accept } H_0 \end{cases}$$

Generalized 0–1 Loss

$$L(\theta, a_0) = \begin{cases} 0 & \text{If } \theta \in \Theta_0 \text{ (correct acceptance)} \\ c_{II} & \text{If } \theta \in \Theta_0^c \text{ (Type II error)} \end{cases}$$

$$L(\theta, a_1) = \begin{cases} c_I & \text{If } \theta \in \Theta_0 \text{ (Type I error)} \\ 0 & \text{If } \theta \in \Theta_0^c \text{ (correct rejection)} \end{cases}$$

c_I: cost of Type I error
c_{II}: cost of Type II error

Risk Function (Expected Loss)

The **risk function** is the expected loss given a decision rule δ

$$R(\theta, \delta) = E_\theta\big[L\big(\theta, \delta(X)\big)\big].$$

Define the **power function** of the test:

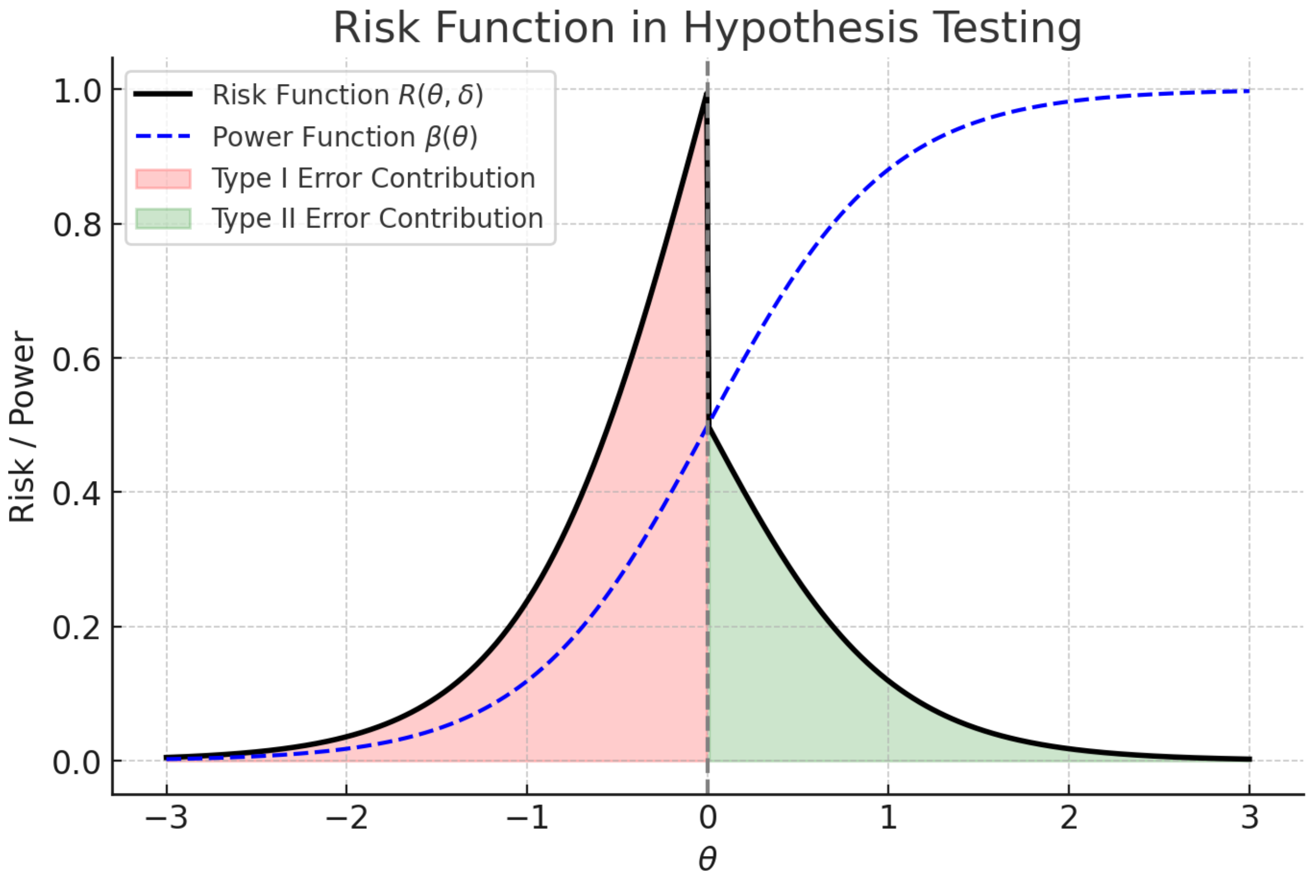
$$\beta(\theta) = P_\theta(\delta(X) = a_1)$$

i.e., the probability of rejecting H_0 under parameter θ.

- If $\theta \in \Theta_0$, $R(\theta, \delta) = c_I \beta(\theta)$.
- If $\theta \in \Theta_0^c$, $R(\theta, \delta) = c_{II}\big(1 - \beta(\theta)\big)$.

These are the probability of a Type I/II error, weighted by its cost.

- The **risk function** formalizes hypothesis testing as a **decision problem** where error costs matter.
- If $c_I = c_{II}$, then this reduces to the simple 0–1 loss (treating both errors equally).
- If $c_I \neq c_{II}$, the decision rule should be chosen to minimize **expected loss**, not just Type I error rate.



- When $\theta \leq 0$ (null true), risk grows with the probability of rejecting H_0 (Type I error)
- When $\theta > 0$ (alternative true), risk decreases with power, since $1 - \beta(\theta)$ is the Type II error probability.

Example:

- Data: $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\sigma = 1, n = 25$.
- Hypotheses: $H_0 : \theta \leq 0$ vs $H_1 : \theta > 0$.
- Test (size $\alpha = 0.05$): reject H_0 when

$$\bar{X} > z_{1-\alpha} \frac{\sigma}{\sqrt{n}} = 1.645 \cdot \frac{1}{5} = 0.329.$$

- Power function:

$$\beta(\theta) = P_\theta(\bar{X} > 0.329) = 1 - \Phi\big(1.645 - \sqrt{n}\,\theta\big) = 1 - \Phi(1.645 - 5\theta).$$

Take $c_I = 2$ (Type I is twice as costly) and $c_{II} = 1$.

Risk is

$$R(\theta) = \begin{cases} c_I\,\beta(\theta), & \theta \leq 0, \\ c_{II}\,[1 - \beta(\theta)], & \theta > 0. \end{cases}$$

- $\theta = 0$ (boundary, null true):

  $\beta(0) = \alpha = 0.05.$

  $R(0) = 2 \times 0.05 = 0.10.$

- $\theta = -0.2$ (null true):

  $\beta(-0.2) = 1 - \Phi(1.645 - 5(-0.2)) = 1 - \Phi(2.645) \approx 0.0041.$

  $R(-0.2) \approx 2 \times 0.0041 = 0.0082.$

- $\theta = 0.2$ (alternative true):

  $\beta(0.2) = 1 - \Phi(1.645 - 1) = 1 - \Phi(0.645) \approx 0.260.$

  Type II probability $= 1 - \beta = 0.740.$

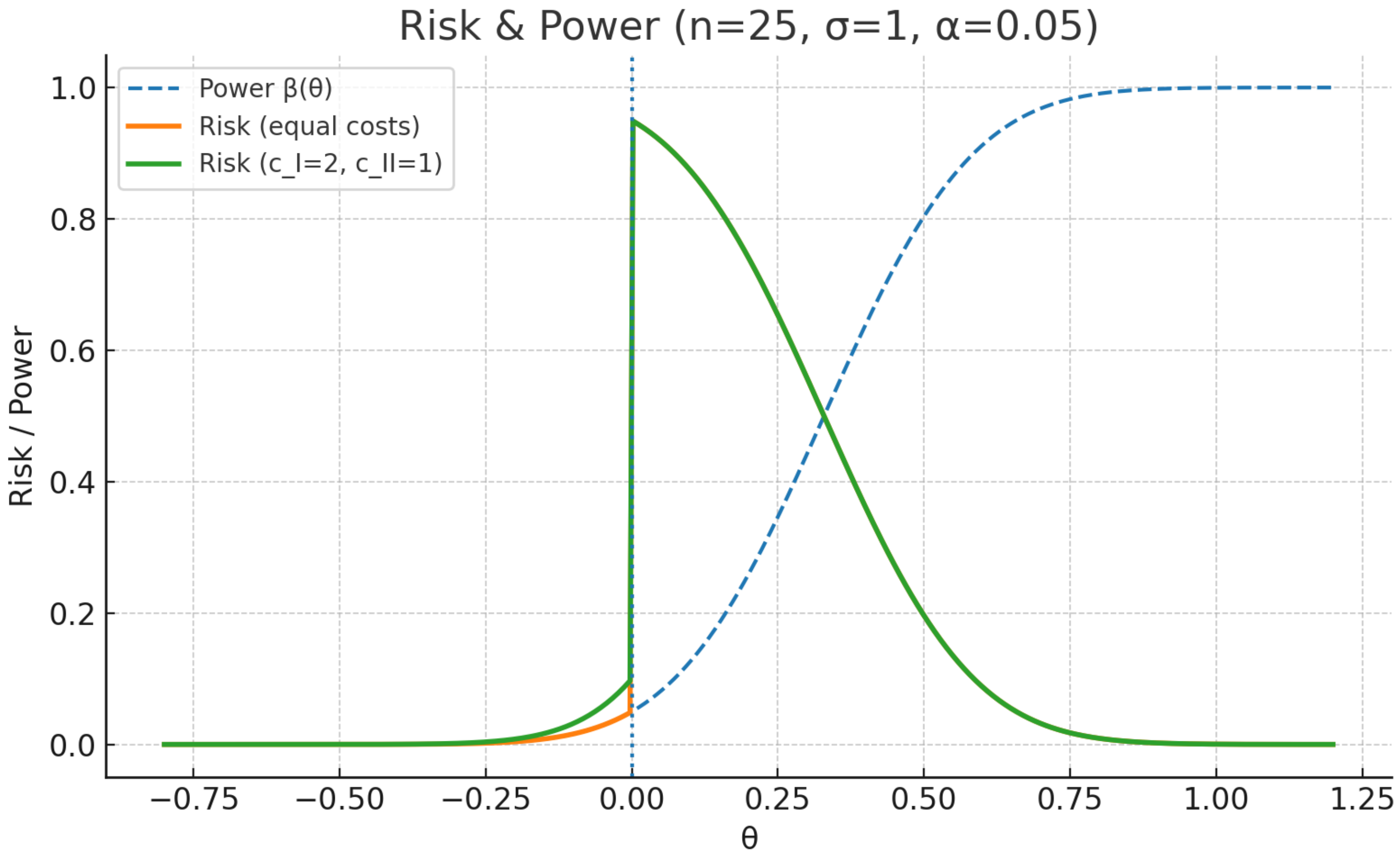
  $R(0.2) = 1 \times 0.740 = 0.740.$

- $\theta = 0.5$ (alternative true):

  $\beta(0.5) = 1 - \Phi(1.645 - 2.5) = 1 - \Phi(-0.855) = \Phi(0.855) \approx 0.803.$

  Type II probability $= 0.197.$

  $R(0.5) \approx 0.197.$


Risk & Power (n=25, σ=1, α=0.05)
Power β(θ)
Risk (equal costs)
Risk (c_I=2, c_II=1)
Risk / Power
1.0
0.8
0.6
0.4
0.2
0.0
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
θ

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition (Chapter 8.3)
- **Book 2. [W]: All of Statistics: Larry Wasserman**
- https://www.probabilitycourse.com/

Online books and courses:

- https://online.stat.psu.edu/stat415/
- https://stat110.hsites.harvard.edu/
- https://bookdown.org/egarpor/inference/

https://en.wikipedia.org/wiki/Misuse_of_p-values