Hypothesis Testing Framework

Statistics · Module 5 — Hypothesis Testing

Node 5.1

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.1

This node establishes the formal language in which all of frequentist hypothesis testing is expressed. It defines tests, errors, power, level, and size as mathematical objects, setting the stage for optimal test construction (Neyman–Pearson, UMP, LR) and calibration (p-values, confidence duality). Every subsequent node in Module 5 inherits the definitions and constraints introduced here.

Formal Definition

Statistical Model

Let

$$X = (X_1, \dots, X_n) \sim P_\theta, \quad \theta \in \Theta \subseteq \mathbb{R}^k,$$

where $\{P_\theta : \theta \in \Theta\}$ is a parametric family dominated by a sigma-finite measure μ, with densities p_θ.

Hypotheses

A **null hypothesis** is a subset $\Theta_0 \subseteq \Theta$. An **alternative hypothesis** is a subset $\Theta_1 \subseteq \Theta$ with $\Theta_0 \cap \Theta_1 = \emptyset$. The testing problem is written

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1.$$

A hypothesis is **simple** if it specifies a unique distribution ($|\Theta_0| = 1$) and **composite** otherwise.

Test Function

A **test function** (or **randomized test**) is a measurable function

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

where $\phi(x)$ is the probability of rejecting H_0 upon observing x. A **non-randomized test** takes values in $\{0, 1\}$ only.

Critical Region

For a non-randomized test, the **critical region** (rejection region) is

$$R = \{x \in \mathcal{X} : \phi(x) = 1\},$$

and the **acceptance region** is R^c.

Intuition

Hypothesis testing formalizes a binary decision under uncertainty: given data, decide whether the evidence is strong enough to reject a default claim H_0. The framework does not ask "Is H_0 true?" in any absolute sense. Instead, it asks: "Under the assumption that H_0 holds, is the observed data sufficiently surprising?"

The test function ϕ is the decision rule. The power function β_ϕ is its operating characteristic, describing how the test behaves across the entire parameter space. The level constraint bounds the worst-case false rejection rate, converting the problem into constrained optimization: maximize power subject to a budget on Type I error.

Structural Role

The test function ϕ and power function β_ϕ encode the complete frequentist decision structure. All subsequent constructions in Module 5 are special cases of this framework:

- Neyman–Pearson (Node 5.2): optimal ϕ for simple-vs-simple problems.
- UMP tests (Node 5.3): optimal ϕ for one-sided composite problems under MLR.
- LR tests (Node 5.4): general-purpose ϕ for composite problems.
- p-values (Node 5.5): continuous calibration of ϕ against H_0.
- Test–confidence duality (Node 5.6): geometric reinterpretation of families of tests.

The framework requires the probabilistic foundations of Module 1, the distributional theory of Module 2, and the estimation concepts of Module 4 (particularly sufficient statistics).

Mathematical Development

Error Types

For a test function ϕ:

Type I error at $\theta \in \Theta_0$:

$$\alpha(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(\text{reject } H_0).$$

Type II error at $\theta \in \Theta_1$:

$$\beta^{\text{II}}(\theta) = 1 - \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(\text{fail to reject } H_0).$$

Power Function

The **power function** of ϕ is

$$\beta_\phi(\theta) = \mathbb{E}_\theta[\phi(X)], \quad \theta \in \Theta.$$

Ideal behavior: $\beta_\phi(\theta)$ small on Θ_0, large on Θ_1.

Level and Size

A test ϕ has **level** α if

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)] \leq \alpha.$$

The **size** of ϕ is

$$\alpha^* = \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)].$$

An exact level α test achieves $\alpha^* = \alpha$.

The Power Function Characterizes the Test

Proposition. Two tests ϕ_1 and ϕ_2 have the same power function if and only if $\phi_1 = \phi_2$ almost everywhere with respect to every P_θ, $\theta \in \Theta$.

Proof. If $\beta_{\phi_1}(\theta) = \beta_{\phi_2}(\theta)$ for all $\theta \in \Theta$, then $\mathbb{E}_\theta[\phi_1(X) - \phi_2(X)] = 0$ for all θ. Since $\phi_1 - \phi_2$ is bounded and $\{P_\theta\}$ is a family of distributions, this means $\int (\phi_1 - \phi_2)\, dP_\theta = 0$ for all θ. If the family $\{P_\theta\}$ is complete (Node 2.6), then $\phi_1 = \phi_2$ a.e. μ. More generally, $\phi_1 = \phi_2$ a.e. with respect to every P_θ. The converse is immediate. $\square$

Unbiased Tests

A test ϕ is **unbiased** at level α if

$$\mathbb{E}_\theta[\phi(X)] \leq \alpha \quad \forall \theta \in \Theta_0, \qquad \mathbb{E}_\theta[\phi(X)] \geq \alpha \quad \forall \theta \in \Theta_1.$$

Unbiasedness ensures the test is at least as likely to reject under the alternative as under the null.

Impossibility of Simultaneous Minimization

Proposition. Unless P_θ is identical for all $\theta \in \Theta_0 \cup \Theta_1$, no test can simultaneously achieve $\beta_\phi(\theta) = 0$ for all $\theta \in \Theta_0$ and $\beta_\phi(\theta) = 1$ for all $\theta \in \Theta_1$ when P_{θ_0} and P_{θ_1} have overlapping support.

Proof. Suppose P_{θ_0} and P_{θ_1} are mutually absolutely continuous for some $\theta_0 \in \Theta_0$, $\theta_1 \in \Theta_1$. If $\phi(x) = 1$ on a set R, then $P_{\theta_0}(R) = 0$ requires R to be a P_{θ_0}-null set. By mutual absolute continuity, R is also P_{θ_1}-null, so $\beta_\phi(\theta_1) = 0$, a contradiction. $\square$

This impossibility drives the entire theory: since perfect separation is unattainable, one must manage the tradeoff between error types.

Worked Examples

Example 1: Normal Mean, Known Variance

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with σ^2 known. Test $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$.

The test statistic is $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$. Under H_0, $Z \sim N(0, 1)$.

The level α non-randomized test is $\phi(x) = \mathbf{1}(Z > z_\alpha)$, where $z_\alpha = \Phi^{-1}(1 - \alpha)$.

The power function is

$$\beta_\phi(\mu) = \mathbb{P}_\mu(Z > z_\alpha) = 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right).$$

At $\mu = \mu_0$: $\beta_\phi(\mu_0) = \alpha$ (size equals level). As $\mu \to \infty$: $\beta_\phi(\mu) \to 1$. As $n \to \infty$ for any fixed $\mu > \mu_0$: $\beta_\phi(\mu) \to 1$ (consistency).

Example 2: Randomized Test for Discrete Data

Let $X \sim \text{Binomial}(10, p)$. Test $H_0 : p = 0.5$ vs $H_1 : p > 0.5$ at level $\alpha = 0.05$.

Under H_0, $\mathbb{P}_{0.5}(X \geq 9) = \binom{10}{9}(0.5)^{10} + (0.5)^{10} = 11/1024 \approx 0.0107$, and $\mathbb{P}_{0.5}(X \geq 8) \approx 0.0547$.

Since $0.0107 < 0.05 < 0.0547$, no non-randomized test achieves exact level 0.05.

The randomized test sets $\phi(x) = 1$ for $x \geq 9$, $\phi(8) = \gamma$, and $\phi(x) = 0$ for $x \leq 7$, where

$$\gamma = \frac{0.05 - 0.0107}{0.0547 - 0.0107} = \frac{0.0393}{0.0440} \approx 0.893.$$

This achieves $\mathbb{E}_{0.5}[\phi(X)] = 0.05$ exactly.

Cross-Links

- **AI:** Type I and Type II errors map directly to false positive rate and false negative rate in binary classification. The ROC curve is the graph of power vs size as the threshold varies.
- **Economics:** Hypothesis tests underpin structural break detection and policy evaluation (e.g., testing whether a treatment effect is zero).
- **Linear Algebra:** The rejection region R is a subset of $\mathbb{R}^n$; in Gaussian models, R is a half-space or cone determined by projections.
- **Quantum:** Quantum hypothesis testing distinguishes quantum states ρ_0 vs ρ_1; the test is a POVM element $0 \leq E \leq I$, generalizing $\phi : \mathcal{X} \to [0, 1]$ to the operator setting.

Structural Compression

Invariant: The fundamental tradeoff: Type I and Type II errors cannot simultaneously be reduced without additional information (larger n, stronger separation of P_{θ_0} and P_{θ_1}).

Mechanism: The level constraint $\sup_{\Theta_0} \mathbb{E}_\theta[\phi] \leq \alpha$ fixes a budget on false rejection; optimality is defined by maximizing power within that budget.

Cross-System Echo: Test functions correspond to decision rules in statistical decision theory; the power function is the risk function under 0-1 loss. In signal detection theory, the same tradeoff appears as the receiver operating characteristic. In machine learning, precision-recall tradeoffs mirror the Type I / Type II structure.

Exercises

1. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. Write the power function of the test that rejects $H_0 : \lambda = \lambda_0$ in favor of $H_1 : \lambda < \lambda_0$ when $\bar{X} > c$, and find c in terms of α and the gamma distribution.
2. Prove that if ϕ is an unbiased level α test, then $\beta_\phi(\theta_0) = \alpha$ for every boundary point θ_0 of Θ_0 that is also in the closure of Θ_1.
3. Consider $X \sim \text{Poisson}(\lambda)$. Construct a randomized test of $H_0 : \lambda = 1$ vs $H_1 : \lambda = 3$ at exact level $\alpha = 0.05$. Compute the power.
4. Prove that for a one-parameter exponential family, the power function $\beta_\phi(\theta)$ of any non-trivial level α test is a continuous function of θ.
5. Let ϕ_1 and ϕ_2 be two level α tests. Show that $\phi_3 = \lambda\phi_1 + (1 - \lambda)\phi_2$ for $\lambda \in [0, 1]$ is also a level α test. What does this say about the set of level α tests?
6. For testing $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ with $X \sim N(\mu, 1)$ (single observation), plot the power function of the test $\phi(x) = \mathbf{1}(|x| > z_{\alpha/2})$ for $\alpha = 0.05$. Show that $\beta_\phi(\mu) \to 1$ as $|\mu| \to \infty$ and that $\beta_\phi(0) = 0.05$.
7. Argue, structurally, why the level constraint $\sup_{\Theta_0} \mathbb{E}_\theta[\phi] \leq \alpha$ is a linear constraint on ϕ, and explain how this convex structure enables the existence of optimal tests via constrained optimization (connecting to the Neyman–Pearson framework of Node 5.2).

Statistics · Module 5 — Hypothesis Testing · Node 5.1
Concepts: hypothesis-testing, probability-space
Prerequisites: 1.3, 2.1, 4.1
Enables: 5.2, 5.3, 5.4, 5.5, 5.6