Module 5 — Hypothesis Testing

Statistics

hbar.university

Hypothesis Testing Framework

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

Neyman–Pearson Theory

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.2

The Neyman–Pearson Lemma is the foundational optimality result in hypothesis testing. It characterizes the most powerful test for simple-vs-simple problems and establishes the likelihood ratio as the fundamental statistic for hypothesis testing. All optimal testing theory (UMP tests, LR tests) descends from this result. It requires the testing framework of Node 5.1 and the distributional theory (densities, likelihood) of Nodes 2.3–2.4.

Formal Definition

Setting

Let P_0 and P_1 be two distributions with densities p_0 and p_1 with respect to a common dominating measure μ. Consider the simple-vs-simple testing problem:

$$H_0 : P = P_0 \quad \text{vs} \quad H_1 : P = P_1.$$

Most Powerful Test

Among all level α tests, a test ϕ^* is **most powerful (MP)** if

$$\mathbb{E}_1[\phi^*(X)] \geq \mathbb{E}_1[\phi(X)]$$

for every test ϕ satisfying $\mathbb{E}_0[\phi(X)] \leq \alpha$.

Intuition

The Neyman–Pearson Lemma answers the question: given a budget α on false alarm probability, how should one allocate rejection probability across the sample space to maximize detection?

The answer is greedy allocation. Rank every point x by how much more likely it is under P_1 than under P_0, i.e., by the likelihood ratio $p_1(x)/p_0(x)$. Reject starting from the points with the highest ratio, spending the budget α on the points that contribute the most power per unit of size.

This is an isoperimetric principle: among all sets of prescribed P_0-measure α, the one with the largest P_1-measure is the upper level set of the likelihood ratio.

Structural Role

The Neyman–Pearson Lemma characterizes optimal tests for simple hypotheses. The likelihood ratio p_1/p_0 is the sufficient statistic for the simple-vs-simple problem: all relevant information for this test is contained in this ratio.

Extensions to composite hypotheses require additional structure: - Monotone likelihood ratio yields UMP tests for one-sided problems (Node 5.3). - Asymptotic approximations yield LR tests for general composite problems (Node 5.4). - The p-value (Node 5.5) calibrates the NP test statistic on the probability scale.

Mathematical Development

Theorem – Neyman–Pearson Lemma

There exists a most powerful level α test of $H_0 : P_0$ vs $H_1 : P_1$, and it has the form

$$\phi^*(x) = \begin{cases} 1 & \text{if } p_1(x) > k\, p_0(x), \\ \gamma & \text{if } p_1(x) = k\, p_0(x), \\ 0 & \text{if } p_1(x) < k\, p_0(x), \end{cases}$$

where $k \geq 0$ and $\gamma \in [0, 1]$ are chosen to satisfy $\mathbb{E}_0[\phi^*(X)] = \alpha$.

Moreover, ϕ^* is unique almost everywhere with respect to μ, except possibly on the set $\{p_1 = k\, p_0\}$.

Proof

Let ϕ be any level α test, i.e., $\mathbb{E}_0[\phi(X)] \leq \alpha$. Define $\Delta(x) = \phi^*(x) - \phi(x)$. We must show $\mathbb{E}_1[\Delta(X)] \geq 0$.

Consider the decomposition of the sample space into three regions:

- On $S_+ = \{x : p_1(x) > k\, p_0(x)\}$: $\phi^*(x) = 1 \geq \phi(x)$, so $\Delta(x) \geq 0$, and $p_1(x) - k\, p_0(x) > 0$.
- On $S_0 = \{x : p_1(x) = k\, p_0(x)\}$: $p_1(x) - k\, p_0(x) = 0$.
- On $S_- = \{x : p_1(x) < k\, p_0(x)\}$: $\phi^*(x) = 0 \leq \phi(x)$, so $\Delta(x) \leq 0$, and $p_1(x) - k\, p_0(x) < 0$.

Therefore $\Delta(x)(p_1(x) - k\, p_0(x)) \geq 0$ everywhere, giving

$$\int \Delta(x)\, p_1(x)\, d\mu(x) \geq k \int \Delta(x)\, p_0(x)\, d\mu(x).$$

The right-hand side is $k(\mathbb{E}_0[\phi^*] - \mathbb{E}_0[\phi]) = k(\alpha - \mathbb{E}_0[\phi]) \geq 0$, since $\mathbb{E}_0[\phi] \leq \alpha$.

Hence $\mathbb{E}_1[\phi^*] \geq \mathbb{E}_1[\phi]$. $\square$

Existence of k and γ

Define $F_0(t) = P_0(p_1(X)/p_0(X) \leq t)$, the CDF of the likelihood ratio under P_0. Set k to be the $(1-\alpha)$-quantile of $p_1(X)/p_0(X)$ under P_0:

$$k = \inf\{t : F_0(t) \geq 1 - \alpha\}.$$

If $P_0(p_1(X)/p_0(X) = k) = 0$, then γ is irrelevant and the test is non-randomized. Otherwise, choose

$$\gamma = \frac{\alpha - P_0(p_1(X)/p_0(X) > k)}{P_0(p_1(X)/p_0(X) = k)}.$$

Corollary: Power Is at Least α

If $P_0 \neq P_1$, then the power of the NP test satisfies $\mathbb{E}_1[\phi^*] \geq \alpha$, with equality only when $p_1 = k\,p_0$ a.e. In particular, the NP test is unbiased.

Proof. From the proof above, $\mathbb{E}_1[\phi^*] \geq k \cdot 0 + \mathbb{E}_1[\phi]$ for the trivial test $\phi \equiv \alpha$, which has $\mathbb{E}_1[\phi] = \alpha$. Since $\int \Delta(x)(p_1(x) - kp_0(x))\,d\mu \geq 0$ with the choice $\phi \equiv \alpha$, and equality requires $p_1 = kp_0$ a.e. on sets where $\Delta \neq 0$. $\square$

Sufficiency of the Likelihood Ratio

The NP test depends on the data only through the likelihood ratio $\Lambda(x) = p_1(x)/p_0(x)$. By the factorization theorem (Node 2.4), $\Lambda(X)$ is a sufficient statistic for the family $\{P_0, P_1\}$. Thus the NP test achieves optimality by reducing the data to a one-dimensional sufficient statistic.

Worked Examples

Example 1: Normal Mean Shift

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$. Test $H_0 : \mu = 0$ vs $H_1 : \mu = 1$ at level α.

The likelihood ratio is

$$\frac{p_1(x)}{p_0(x)} = \prod_{i=1}^{n} \frac{e^{-(x_i-1)^2/2}}{e^{-x_i^2/2}} = \exp\left(\sum_{i=1}^{n} x_i - \frac{n}{2}\right).$$

This is monotone increasing in $\bar{x} = n^{-1}\sum x_i$. The NP test rejects when $\bar{X} > c$, where c satisfies $P_0(\bar{X} > c) = \alpha$.

Under H_0, $\bar{X} \sim N(0, 1/n)$, so $c = z_\alpha/\sqrt{n}$.

The power is

$$\beta(\mu = 1) = P_1(\bar{X} > z_\alpha/\sqrt{n}) = 1 - \Phi(z_\alpha - \sqrt{n}).$$

For $n = 25$, $\alpha = 0.05$: $c = 1.645/5 = 0.329$, and power $= 1 - \Phi(1.645 - 5) = 1 - \Phi(-3.355) \approx 0.9996$.

Example 2: Exponential Rate

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$ with density $\lambda e^{-\lambda x}$. Test $H_0 : \lambda = 1$ vs $H_1 : \lambda = 2$ at level $\alpha = 0.05$.

The likelihood ratio is

$$\frac{p_1(x)}{p_0(x)} = \frac{2^n e^{-2\sum x_i}}{e^{-\sum x_i}} = 2^n \exp\left(-\sum_{i=1}^{n} x_i\right).$$

This is decreasing in $T = \sum x_i$, so the NP test rejects for small T.

Under H_0, $2T = 2\sum X_i \sim \chi^2_{2n}$. For $n = 5$, reject when $2T < \chi^2_{10,0.95}$. From tables, $\chi^2_{10,0.95} = 3.940$, so reject when $T < 1.970$.

Power: Under H_1, $4T = 4\sum X_i \sim \chi^2_{10}$, so $P_1(T < 1.970) = P(\chi^2_{10} < 7.880) \approx 0.36$.

Cross-Links

- **AI:** The likelihood ratio is the Bayes-optimal classifier under equal priors and 0-1 loss. The NP lemma formalizes the optimality of log-odds thresholding in binary classification.
- **Economics:** In mechanism design, the NP structure appears in optimal auction theory (Myerson): allocate the good to the bidder with the highest virtual value, analogous to rejecting for the highest likelihood ratio.
- **Linear Algebra:** In Gaussian models, the NP critical region is a half-space in $\mathbb{R}^n$, determined by the projection of the data onto the direction $\mu_1 - \mu_0$.
- **Quantum:** The Helstrom bound on quantum state discrimination is the quantum analogue of the NP bound: the optimal POVM element projects onto the positive part of $\rho_1 - k\rho_0$.

Structural Compression

Invariant: Among level α tests of P_0 vs P_1, the likelihood ratio test is most powerful.

Mechanism: Reject when $p_1(x)/p_0(x) > k$: the decision boundary is a level set of the likelihood ratio, allocating rejection probability greedily to the highest-evidence points.

Cross-System Echo: The NP lemma is an instance of the general isoperimetric principle: among all sets of fixed measure under one distribution, the one with maximal measure under another is the upper level set of the Radon–Nikodym derivative. This principle recurs in information theory (channel coding), optimal transport (Monge sets), and quantum information (Helstrom measurement).

Exercises

1. Prove that the Neyman–Pearson test is unique almost everywhere when the distribution of $p_1(X)/p_0(X)$ under P_0 is continuous.
2. Show that the power $\mathbb{E}_1[\phi^*(X)]$ is a non-decreasing function of α, and strictly increasing when P_0 and P_1 are mutually absolutely continuous.
3. Let $X \sim \text{Poisson}(\lambda)$. Derive the NP test of $H_0 : \lambda = 1$ vs $H_1 : \lambda = 3$ at level $\alpha = 0.05$. Compute the threshold k, the randomization probability γ, and the exact power.
4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with σ^2 known. Show that the NP test of $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1$ (with $\mu_1 > \mu_0$) rejects for large $\bar{X}$, and that the critical value does not depend on μ_1.
5. Prove that if ϕ^* is the NP test at level α and ϕ^* has power β^*, then the NP test of $H_0 : P_1$ vs $H_1 : P_0$ at level $1 - \beta^*$ has power $1 - \alpha$.
6. Let $P_0 = N(0, 1)$ and $P_1 = N(0, \sigma^2)$ with $\sigma^2 > 1$. Show that the NP test rejects for large $|X|$, and compute the threshold.
7. Argue, structurally, why the Neyman–Pearson Lemma is an instance of Lagrangian optimization: identify the objective, the constraint, and the role of the multiplier k, and explain why the complementary slackness condition corresponds to the randomization on the boundary $\{p_1 = kp_0\}$.

Uniformly Most Powerful Tests

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.3

This node extends the Neyman–Pearson optimality result from simple to composite hypotheses. When the parametric family possesses a monotone likelihood ratio, a single test can be simultaneously most powerful against every point alternative on one side, yielding a uniformly most powerful (UMP) test. The key structural result is the Karlin–Rubin theorem, which reduces composite one-sided testing to the simple-vs-simple NP framework via the MLR property. This node requires Node 5.2 (NP Lemma) and the exponential family theory of Node 2.6.

Formal Definition

UMP Test

Consider a composite testing problem:

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

A level α test ϕ^* is **uniformly most powerful (UMP)** if

$$\mathbb{E}_\theta[\phi^*(X)] \geq \mathbb{E}_\theta[\phi(X)] \quad \forall \theta \in \Theta_1$$

for every level α test ϕ.

A UMP test is simultaneously most powerful against every point in Θ_1.

Monotone Likelihood Ratio

A family $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ with densities $\{p_\theta\}$ has **monotone likelihood ratio (MLR)** in a statistic $T(X)$ if for every $\theta_1 < \theta_2$, the ratio

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}$$

is a non-decreasing function of $T(x)$ (on the set where $p_{\theta_1}(x) > 0$).

Intuition

The NP Lemma gives the best test for a single point alternative. The challenge with composite alternatives is that different point alternatives might require different critical regions.

The MLR property resolves this conflict: it guarantees that the likelihood ratio between any two parameter values is monotone in the same statistic T. Consequently, the NP test for every pair (θ_0, θ_1) with $\theta_1 > \theta_0$ takes the same form — reject when T is large. A single threshold test on T is therefore optimal against the entire composite alternative simultaneously.

The MLR condition is a coherence requirement: as θ increases, the distribution of T shifts in a consistent direction.

Structural Role

UMP tests occupy a narrow but important niche: they exist for one-sided hypotheses in MLR families, which includes all one-parameter exponential families with increasing natural parameter.

For two-sided alternatives ($H_1 : \theta \neq \theta_0$), UMP tests generally do not exist. This non-existence motivates the shift to: - Unbiased UMP (UMPU) tests for two-sided problems in exponential families. - Likelihood ratio tests (Node 5.4) as a general-purpose fallback.

The Karlin–Rubin theorem connects the exponential family structure (Node 2.6) to the testing framework (Node 5.1) through the NP optimality criterion (Node 5.2).

Mathematical Development

Theorem – Karlin–Rubin

Suppose the family $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ has MLR in $T(X)$. For testing

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0,$$

the test

$$\phi^*(x) = \begin{cases} 1 & T(x) > c, \\ \gamma & T(x) = c, \\ 0 & T(x) < c, \end{cases}$$

where c and γ are chosen so that $\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha$, is UMP level α.

Proof

Step 1: Most powerful at each $\theta_1 > \theta_0$.

Fix any $\theta_1 > \theta_0$. By the Neyman–Pearson Lemma, the most powerful level α test of $H_0' : \theta = \theta_0$ vs $H_1' : \theta = \theta_1$ rejects for large values of

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

By the MLR property, this ratio is non-decreasing in $T(x)$. Therefore the NP critical region $\{p_{\theta_1}/p_{\theta_0} > k\}$ is equivalent to $\{T(x) > c'\}$ for some threshold c'.

The level constraint $\mathbb{E}_{\theta_0}[\phi] = \alpha$ uniquely determines the threshold (and randomization constant) when the distribution of T under θ_0 is specified. Since this threshold depends only on θ_0 and α, not on θ_1, the same test ϕ^* is most powerful against every $\theta_1 > \theta_0$.

Step 2: Level α over all of Θ_0.

We must verify $\mathbb{E}_\theta[\phi^*(X)] \leq \alpha$ for all $\theta \leq \theta_0$, not just at θ_0.

The power function $\beta(\theta) = \mathbb{E}_\theta[\phi^*(X)]$ is non-decreasing (proved below). Since $\beta(\theta_0) = \alpha$, we have $\beta(\theta) \leq \alpha$ for all $\theta \leq \theta_0$. $\square$

Proposition: Monotonicity of the Power Function

If $\{P_\theta\}$ has MLR in T and ϕ^* is a threshold test on T, then $\beta(\theta) = \mathbb{E}_\theta[\phi^*(X)]$ is non-decreasing in θ.

Proof. Let $\theta_1 < \theta_2$. The test ϕ^* is the NP test of P_{θ_1} vs P_{θ_2} at some level $\beta(\theta_1)$. By the NP corollary (the NP test is unbiased), $\beta(\theta_2) \geq \beta(\theta_1)$. □

MLR in Exponential Families

For a one-parameter exponential family with canonical density

$$p_\theta(x) = h(x) \exp\{\eta(\theta)T(x) - A(\theta)\},$$

if $\eta(\theta)$ is strictly increasing in θ, then for $\theta_1 < \theta_2$:

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = \exp\{[\eta(\theta_2) - \eta(\theta_1)]T(x) - [A(\theta_2) - A(\theta_1)]\},$$

which is strictly increasing in $T(x)$ since $\eta(\theta_2) - \eta(\theta_1) > 0$. Thus exponential families with increasing natural parameter have MLR in the natural sufficient statistic T.

Non-Existence of UMP for Two-Sided Alternatives

Proposition. For a one-parameter exponential family, there is no UMP level α test of $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ (when $0 < \alpha < 1$).

Proof sketch. The UMP test of $\theta = \theta_0$ vs $\theta > \theta_0$ rejects for large T. The UMP test of $\theta = \theta_0$ vs $\theta < \theta_0$ rejects for small T. These are different tests, so no single test is simultaneously most powerful against alternatives on both sides. □

Unbiased UMP (UMPU) Tests

For two-sided problems in exponential families, one restricts to unbiased tests and seeks the UMP within that class. A level α test is **unbiased** if $\beta(\theta) \geq \alpha$ for all $\theta \in \Theta_1$.

For testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ in a one-parameter exponential family, the UMPU test rejects when $T \leq c_1$ or $T \geq c_2$, with randomization at the boundaries, where c_1, c_2 are determined by

$$\mathbb{E}_{\theta_0}[\phi(X)] = \alpha, \qquad \mathbb{E}_{\theta_0}[T(X)\phi(X)] = \alpha\, \mathbb{E}_{\theta_0}[T(X)].$$

Worked Examples

Example 1: Normal Mean (Known Variance)

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

This is a one-parameter exponential family with $T = \sum X_i$ and natural parameter $\eta = \mu/\sigma^2$, which is increasing in μ. By the Karlin–Rubin theorem, the UMP level α test rejects when $\bar{X} > c$.

Under μ_0: $\bar{X} \sim N(\mu_0, \sigma^2/n)$, so $c = \mu_0 + z_\alpha \sigma/\sqrt{n}$.

Power function:

$$\beta(\mu) = 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right).$$

This is strictly increasing in μ, confirming monotonicity.

Example 2: Bernoulli Proportion

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. Test $H_0 : p \leq 1/2$ vs $H_1 : p > 1/2$ at level $\alpha = 0.05$, with $n = 20$.

The sufficient statistic is $T = \sum X_i \sim \text{Binomial}(20, p)$. The Bernoulli family has MLR in T since $\eta(p) = \log(p/(1-p))$ is strictly increasing.

Under $p = 1/2$: - $P_{1/2}(T \geq 15) = \sum_{k=15}^{20} \binom{20}{k}(1/2)^{20} \approx 0.0207$. - $P_{1/2}(T \geq 14) \approx 0.0577$.

Non-randomized: reject when $T \geq 15$, with size $0.0207 < 0.05$.

Randomized UMP test: $\phi(x) = 1$ for $T \geq 15$, $\phi(x) = \gamma$ for $T = 14$, where

$$\gamma = \frac{0.05 - 0.0207}{P_{1/2}(T = 14)} = \frac{0.0293}{0.0370} \approx 0.792.$$

Power at $p = 0.7$: $\beta(0.7) = P_{0.7}(T \geq 15) + 0.792 \cdot P_{0.7}(T = 14) \approx 0.584 + 0.792 \cdot 0.192 \approx 0.736$.

Cross-Links

- **AI:** Monotone decision boundaries in UMP tests correspond to threshold classifiers on a score function. In calibrated binary classifiers under log-loss, the optimal decision rule is monotone in the predicted probability.
- **Economics:** One-sided tests arise in policy evaluation (e.g., testing whether a program has a positive effect). The UMP property ensures no other level-α test has higher power to detect the posited direction of effect.
- **Linear Algebra:** The sufficient statistic T is a linear functional of the data in exponential families; the UMP critical region is a half-space in the sufficient statistic space.
- **Quantum:** For quantum state families with a monotone structure (e.g., thermal states parameterized by temperature), the optimal measurement for one-sided discrimination is the quantum analogue of the UMP test.

Structural Compression

Invariant: MLR reduces the composite testing problem to a sequence of simple-vs-simple problems all solved by the same critical region.

Mechanism: The statistic T orders the family monotonically; the threshold test on T is simultaneously Neyman–Pearson optimal against every point alternative in H_1, because the NP critical region is the same for every $\theta_1 \in \Theta_1$.

Cross-System Echo: The monotone likelihood ratio is a stochastic ordering condition: T under P_{θ_2} is stochastically larger than under P_{θ_1} when $\theta_2 > \theta_1$. This ordering principle recurs in reliability theory (increasing failure rate families), decision theory (monotone decision procedures), and auction theory (revenue monotonicity under regular distributions).

Exercises

1. Show that the Bernoulli(θ) family has MLR in $T = \sum_{i=1}^{n} X_i$, and derive the UMP test for $H_0 : \theta \leq 1/2$ vs $H_1 : \theta > 1/2$ at level $\alpha = 0.10$ with $n = 10$.
2. Prove that the power function of the Karlin–Rubin test is non-decreasing in θ, using only the NP Lemma and the MLR property.

3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Verify the MLR property in $T = \sum X_i$ and find the UMP test of $H_0 : \lambda \leq 1$ vs $H_1 : \lambda > 1$ for $n = 10$ and $\alpha = 0.05$.

4. Prove that for a one-parameter exponential family with strictly increasing $\eta(\theta)$, there is no UMP test of $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

5. For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_0, \beta)$ with α_0 known, show that the family has MLR in $\sum X_i$ and state the UMP test for $H_0 : \beta \leq \beta_0$ vs $H_1 : \beta > \beta_0$.

6. Construct the UMPU test of $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ for $X \sim N(\mu, 1)$ (single observation). Verify that it coincides with the two-sided z-test.

7. Argue, structurally, why the MLR property is the precise condition under which the NP lemma's pointwise optimality lifts to uniform optimality, and explain what breaks for families that fail to satisfy MLR (connecting to the non-existence result for two-sided alternatives).

Likelihood Ratio Tests

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.4

The likelihood ratio (LR) test is the general-purpose construction for composite hypothesis testing. Where Neyman–Pearson handles simple-vs-simple and UMP handles one-sided MLR families, LR tests apply to arbitrary parametric constraints without requiring special structural conditions. Wilks' theorem provides the asymptotic reference distribution, making LR tests practically computable in large samples. This node requires the testing framework (Node 5.1), density and likelihood theory (Node 2.3), and MLE theory (Node 3.2).

Formal Definition

Generalized Likelihood Ratio Statistic

Let $X_1, \ldots, X_n \sim p_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$. Consider

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta \setminus \Theta_0,$$

where $\Theta_0 \subset \Theta$ is defined by r smooth equality constraints, so $\dim(\Theta_0) = k - r$.

The **generalized likelihood ratio statistic** is

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_0} L(\theta; X)}{\sup_{\theta \in \Theta} L(\theta; X)} = \frac{L(\hat{\theta}_0; X)}{L(\hat{\theta}_n; X)},$$

where $\hat{\theta}_n$ is the unrestricted MLE and $\hat{\theta}_0$ is the MLE restricted to Θ_0.

Note $\Lambda_n \in (0, 1]$; the test rejects H_0 for small values of Λ_n.

Log-Likelihood Form

The standard form is

$$-2 \log \Lambda_n = 2 \left[\ell(\hat{\theta}_n) - \ell(\hat{\theta}_0) \right],$$

which is non-negative. The test rejects H_0 for large values of $-2 \log \Lambda_n$.

Intuition

The LR statistic measures how much the data favors the unrestricted model over the restricted model. If the constraint $\theta \in \Theta_0$ costs little in likelihood, Λ_n is near 1 and there is no evidence against H_0. If the constraint forces a substantial loss in fit, Λ_n is small and the data contradicts H_0.

The factor of -2 in the log transform is not arbitrary: it is chosen so that the asymptotic distribution under H_0 is exactly chi-squared, without additional scaling. This normalization arises from the quadratic approximation to the log-likelihood near its maximum, with the Fisher information providing the natural metric.

Structural Role

LR tests provide a universal construction for composite hypotheses. Their advantages:

1. **Generality.** Applicable to any parametric model with smooth constraints.
2. **Invariance.** The LR statistic is invariant under reparameterization of θ.
3. **Asymptotic universality.** Wilks' theorem gives a single reference distribution (χ^2_r) regardless of the specific model.

The LR test, Wald test, and Score (Rao) test form the "holy trinity" of asymptotic tests (Node 4.4). All three converge to the same first-order asymptotic limit, but they differ in finite samples and in computational convenience.

Mathematical Development

Theorem – Wilks (1938)

Regularity conditions. Assume: (i) Θ is an open subset of $\mathbb{R}^k$; (ii) $\Theta_0 = \{\theta \in \Theta : g(\theta) = 0\}$ where $g : \mathbb{R}^k \to \mathbb{R}^r$ is continuously differentiable with rank r Jacobian; (iii) the log-likelihood is three times continuously differentiable with non-singular Fisher information $I(\theta)$; (iv) the MLE $\hat{\theta}_n$ is consistent and asymptotically normal.

Statement. Under H_0 (with true parameter $\theta_0 \in \Theta_0$),

$$-2\log\Lambda_n \xrightarrow{d} \chi^2_r \quad \text{as } n \to \infty,$$

where $r = \dim(\Theta) - \dim(\Theta_0)$.

Proof Sketch

Step 1: Taylor expansion. Expand $\ell(\theta)$ around $\hat{\theta}_n$ to second order:

$$\ell(\theta) \approx \ell(\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)^\top \left[-\nabla^2 \ell(\hat{\theta}_n)\right](\theta - \hat{\theta}_n).$$

Define the observed information $\hat{J}_n = -\nabla^2\ell(\hat{\theta}_n)/n$. By the law of large numbers, $\hat{J}_n \xrightarrow{p} I(\theta_0)$.

Step 2: Quadratic approximation. Using the expansion at the restricted MLE:

$$-2\log\Lambda_n = 2[\ell(\hat{\theta}_n) - \ell(\hat{\theta}_0)] \approx n(\hat{\theta}_n - \hat{\theta}_0)^\top \hat{J}_n (\hat{\theta}_n - \hat{\theta}_0).$$

Step 3: Asymptotic normality. Under H_0, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$.

Step 4: Constrained projection. The restricted MLE $\hat{\theta}_0$ is the projection of $\hat{\theta}_n$ onto Θ_0 in the metric defined by $I(\theta_0)$. The difference $\hat{\theta}_n - \hat{\theta}_0$ lies in the r-dimensional normal space to Θ_0 at θ_0.

Step 5: Chi-squared conclusion. The quadratic form $n(\hat{\theta}_n - \hat{\theta}_0)^\top I(\theta_0)(\hat{\theta}_n - \hat{\theta}_0)$ is asymptotically a sum of squares of r independent standard normals, hence χ^2_r. □

Level α LR Test

Reject H_0 when

$$-2\log\Lambda_n > \chi^2_{r,\alpha},$$

where $\chi^2_{r,\alpha}$ is the upper α quantile of χ^2_r. This test has asymptotic level α by Wilks' theorem.

Relationship to Wald and Score Tests

Under the same regularity conditions:

Wald statistic:

$$W_n = n(\hat{\theta}_n - \theta_0)^\top \hat{I}_n(\hat{\theta}_n - \theta_0) \xrightarrow{d} \chi^2_r.$$

Score (Rao) statistic:

$$S_n = \frac{1}{n}\nabla\ell(\hat{\theta}_0)^\top I(\hat{\theta}_0)^{-1}\nabla\ell(\hat{\theta}_0) \xrightarrow{d} \chi^2_r.$$

All three statistics satisfy $W_n = -2\log\Lambda_n + o_p(1) = S_n + o_p(1)$ under H_0, so they are first-order asymptotically equivalent.

Bartlett Correction

The LR statistic admits a Bartlett correction: there exists a constant c (depending on the model) such that

$$(1 - c/n)\cdot(-2\log\Lambda_n)$$

has a χ^2_r distribution accurate to order $O(n^{-2})$, improving finite-sample performance. This correction is specific to the LR statistic and does not apply to the Wald or Score statistics.

Worked Examples

Example 1: Testing a Normal Mean (Known Variance)

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$. Test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.

Here $k = 1$, $r = 1$, $\hat{\mu}_n = \bar{X}$, $\hat{\mu}_0 = \mu_0$.

$$\ell(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum(X_i - \mu)^2.$$

$$-2\log\Lambda_n = 2[\ell(\bar{X}) - \ell(\mu_0)] = n(\bar{X} - \mu_0)^2.$$

Under H_0, $\sqrt{n}(\bar{X} - \mu_0) \sim N(0,1)$, so $-2\log\Lambda_n \sim \chi^2_1$ exactly.

Reject when $n(\bar{X} - \mu_0)^2 > \chi^2_{1,\alpha} = z^2_{\alpha/2}$, which is equivalent to the two-sided z-test $|\bar{X} - \mu_0| > z_{\alpha/2}/\sqrt{n}$.

Example 2: Testing Equality of Two Normal Means

Let $X_1, \ldots, X_m \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_2, \sigma^2)$, σ^2 known. Test $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.

Parameters: $\theta = (\mu_1, \mu_2) \in \mathbb{R}^2$, constraint $g(\theta) = \mu_1 - \mu_2 = 0$, so $r = 1$.

Unrestricted MLEs: $\hat{\mu}_1 = \bar{X}$, $\hat{\mu}_2 = \bar{Y}$. Restricted MLE: $\hat{\mu}_0 = (m\bar{X} + n\bar{Y})/(m+n)$.

$$-2\log\Lambda_n = \frac{mn}{(m+n)\sigma^2}(\bar{X} - \bar{Y})^2.$$

Under H_0, $\bar{X} - \bar{Y} \sim N(0, \sigma^2(1/m + 1/n))$, so $-2\log\Lambda_n \sim \chi^2_1$ exactly.

Example 3: Testing Independence in a Multinomial

Consider a 2×2 contingency table with cell counts $(n_{11}, n_{12}, n_{21}, n_{22})$ from a multinomial with $N = \sum n_{ij}$ and cell probabilities p_{ij}. Test $H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$ (independence).

Here $k = 3$ (full multinomial has 3 free parameters), $\dim(\Theta_0) = 2$ (row and column marginals determine Θ_0), so $r = 1$.

The LR statistic is

$$-2\log\Lambda_n = 2\sum_{i,j} n_{ij}\log\frac{n_{ij}}{\hat{e}_{ij}},$$

where $\hat{e}_{ij} = n_{i\cdot}n_{\cdot j}/N$ are the expected counts under independence.

Under H_0, $-2\log\Lambda_n \xrightarrow{d} \chi^2_1$ as $N \to \infty$.

Cross-Links

- **AI:** Model comparison in machine learning (e.g., nested neural network architectures, feature selection) uses likelihood ratio principles. The deviance in generalized linear models is exactly $-2\log\Lambda_n$.
- **Economics:** LR tests are standard for testing restrictions in econometric models (e.g., testing coefficient restrictions in regression, testing overidentifying restrictions in GMM).
- **Linear Algebra:** The quadratic form $n(\hat{\theta}_n - \hat{\theta}_0)^\top I(\theta_0)(\hat{\theta}_n - \hat{\theta}_0)$ is a Mahalanobis distance; its chi-squared distribution follows from the spectral decomposition of the information matrix.
- **Quantum:** Quantum likelihood ratio tests for composite hypotheses on quantum states follow the same structure; the quantum Fisher information matrix replaces the classical one, and the SLD (symmetric logarithmic derivative) plays the role of the score.

Structural Compression

Invariant: The LR statistic compresses all information about the constraint $\theta \in \Theta_0$ into a single scalar via the log-likelihood difference.

Mechanism: Second-order Taylor expansion of the log-likelihood under regularity reduces $-2\log\Lambda_n$ to a chi-squared quadratic form in r asymptotically normal directions, with r equal to the codimension of the constraint.

Cross-System Echo: The chi-squared reference distribution is shared by Wald and Score statistics; all three converge to the same first-order asymptotic limit. The $-2\log\Lambda$ form reappears in information criteria (AIC $= -2\ell + 2k$ penalizes by model dimension), in deviance analysis for GLMs, and in the quantum Chernoff bound for asymptotic state discrimination.

Exercises

1. Derive the LR statistic for testing $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$ in the model $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with μ known. Verify the asymptotic chi-squared distribution.

2. Show that the LR statistic is invariant under reparameterization: if $\eta = g(\theta)$ is a diffeomorphism, the LR statistic computed in the η-parameterization equals the one in the θ-parameterization.

3. For the multinomial goodness-of-fit test ($H_0 : p = p_0$ with p_0 fully specified), compute $-2 \log \Lambda_n$ and verify that $r = k - 1$ where k is the number of categories.

4. Prove that for a one-parameter model with simple null $H_0 : \theta = \theta_0$, the LR statistic equals the square of the signed likelihood root: $-2 \log \Lambda_n = [\text{sign}(\hat{\theta}_n - \theta_0)\sqrt{-2 \log \Lambda_n}]^2$, and the signed root is asymptotically $N(0, 1)$.

5. Compute the LR test for $H_0 : \lambda_1 = \lambda_2$ vs $H_1 : \lambda_1 \neq \lambda_2$ given independent samples $X_1, \ldots, X_m \overset{\text{iid}}{\sim} \text{Poisson}(\lambda_1)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda_2)$.

6. State the Bartlett correction for the LR test of $H_0 : \mu = 0$ in a $N(\mu, 1)$ model and verify that the corrected statistic has improved distributional accuracy.

7. Argue, structurally, why the LR test is the natural successor to the NP lemma for composite hypotheses: identify what structural property is lost when moving from simple to composite testing, and explain how the asymptotic chi-squared approximation substitutes for the exact NP optimality.

p-Values

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.5

The p-value is the continuous calibration of a test statistic against the null distribution. It encodes an entire family of hypothesis tests into a single random variable, connecting the binary decision framework of Node 5.1 to the continuous probability scale. This node requires the testing framework (Node 5.1) and the NP theory (Node 5.2) that motivates specific test statistics. It directly enables the test–confidence duality (Node 5.6) through the relationship between p-values and confidence set membership.

Formal Definition

p-Value

Let $T(X)$ be a test statistic for $H_0 : \theta \in \Theta_0$, with large values of T constituting evidence against H_0. The **p-value** is the statistic

$$p(X) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\big(T(X') \geq T(X)\big),$$

where $X' \sim P_\theta$ is an independent copy drawn under H_0.

The p-value is the probability, under the least favorable null distribution, of observing a test statistic at least as extreme as the observed value.

Valid p-Value

A statistic $p(X)$ taking values in $[0, 1]$ is a **valid p-value** for H_0 if

$$\mathbb{P}_\theta\big(p(X) \leq u\big) \leq u \quad \forall u \in [0, 1],\ \forall \theta \in \Theta_0.$$

This is precisely the condition that $p(X)$ is stochastically no smaller than Uniform$(0, 1)$ under every null distribution.

Intuition

The p-value answers: "If the null hypothesis were true, how surprising is the observed data?" A small p-value means the observed test statistic is in the tail of the null distribution, providing evidence against H_0.

Crucially, the p-value is NOT: - The probability that H_0 is true. - The probability that the data arose by chance. - A measure of effect size or practical significance.

The p-value is a frequentist object: it measures the extremity of the observed statistic relative to the null reference distribution. Its uniform distribution under H_0 is the key structural property that makes it a valid calibration device.

Structural Role

The p-value occupies a central position in applied statistics as the standard summary of evidence against H_0. It mediates between:

- The test statistic $T(X)$, which depends on the specific model and test construction.
- The binary decision (reject/not reject), which depends on the chosen level α.

By reporting the p-value rather than just the binary decision, the analyst communicates the strength of evidence on a universal scale.

The p-value connects directly to confidence sets (Node 5.6): the set of parameter values for which the p-value exceeds α forms a $(1-\alpha)$ confidence set.

Mathematical Development

Uniformity Under Simple Null

Theorem. If $H_0 : P = P_0$ is simple and $T(X)$ has a continuous distribution under P_0, then

$$p(X) = 1 - F_0(T(X)) \sim \text{Uniform}(0,1) \quad \text{under } P_0,$$

where F_0 is the CDF of T under P_0.

Proof. By the probability integral transform, if T has continuous CDF F_0, then $U = F_0(T) \sim \text{Uniform}(0,1)$ under P_0. Therefore $p(X) = 1 - U \sim \text{Uniform}(0,1)$. □

Validity Under Composite Null

Theorem. For a composite null, the p-value defined via the supremum is valid:

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(p(X) \leq u) \leq u \quad \forall u \in [0,1].$$

Proof. For any $\theta \in \Theta_0$:

$$\mathbb{P}_\theta(p(X) \leq u) = \mathbb{P}_\theta\left(\sup_{\theta' \in \Theta_0} \mathbb{P}_{\theta'}(T(X') \geq T(X)) \leq u\right) \leq \mathbb{P}_\theta(\mathbb{P}_\theta(T(X') \geq T(X)) \leq u)\,.$$

If T has a continuous distribution under θ, the inner probability is $1 - F_\theta(T(X))$, and by the probability integral transform, $\mathbb{P}_\theta(1 - F_\theta(T(X)) \leq u) = u$. For non-continuous T, the inequality $\leq u$ holds by the properties of CDFs. □

Connection to Level α Tests

Proposition. Rejecting H_0 when $p(X) \leq \alpha$ defines a level α test for every $\alpha \in (0,1]$ simultaneously.

Proof. By validity, $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(p(X) \leq \alpha) \leq \alpha$. □

The p-value therefore carries strictly more information than any single fixed-α binary decision: it encodes the entire family of level-α tests as α ranges over $(0,1]$.

Relationship to Confidence Sets

For each $\theta_0 \in \Theta$, let $p_{\theta_0}(X)$ denote the p-value for testing $H_0 : \theta = \theta_0$. Define

$$C_\alpha(X) = \{\theta_0 \in \Theta : p_{\theta_0}(X) > \alpha\}.$$

Then $C_\alpha(X)$ is a $(1-\alpha)$ confidence set. This is the test inversion construction formalized in Node 5.6.

Lindley's Paradox (Bayesian Interpretation)

The frequentist p-value can conflict sharply with Bayesian posterior probabilities.

Setup. Test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ with $X \sim N(\mu, 1/n)$. Suppose the prior assigns probability $\pi_0 = 1/2$ to H_0 and, under H_1, $\mu \sim N(0, \tau^2)$.

The Bayes factor in favor of H_0 is

$$\mathrm{BF}_{01} = \sqrt{1 + n\tau^2} \cdot \exp\left(-\frac{n\bar{x}^2}{2} \cdot \frac{n\tau^2}{1 + n\tau^2}\right).$$

For fixed $\bar{x} = z_\alpha/\sqrt{n}$ (i.e., the p-value is exactly α), as $n \to \infty$: the frequentist p-value remains constant at α, but $\mathrm{BF}_{01} \to \infty$, meaning the Bayesian posterior increasingly favors H_0.

This is **Lindley's paradox**: a fixed p-value becomes increasingly compatible with H_0 from the Bayesian perspective as sample size grows. The paradox arises because the p-value calibrates against the null alone, while the Bayes factor accounts for the prior predictive distribution under both hypotheses.

Multiple Testing

When m hypotheses $H_{0,1}, \ldots, H_{0,m}$ are tested simultaneously with p-values $p_1, \ldots, p_m$, the probability of at least one false rejection (familywise error rate, FWER) can far exceed α.

Bonferroni correction. Reject $H_{0,i}$ if $p_i \leq \alpha/m$. Then FWER $\leq \alpha$ by the union bound. This is conservative: it controls FWER but sacrifices power.

Benjamini–Hochberg (BH) procedure. Controls the **false discovery rate (FDR)** $= \mathbb{E}[V/R]$ where V = number of false rejections, R = total rejections ($V/R := 0$ if $R = 0$).

Algorithm: 1. Order p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$. 2. Find the largest k such that $p_{(k)} \leq k\alpha/m$. 3. Reject all $H_{0,(i)}$ for $i \leq k$.

Theorem (Benjamini–Hochberg, 1995). If the p-values corresponding to the true nulls are independent, then the BH procedure controls FDR $\leq \alpha \cdot m_0/m \leq \alpha$, where m_0 is the number of true nulls.

Worked Examples

Example 1: z-Test p-Value

Let $X_1, \ldots, X_{25} \overset{\text{iid}}{\sim} N(\mu, 4)$. Test $H_0 : \mu = 10$ vs $H_1 : \mu > 10$. Observed $\bar{x} = 10.92$.

Test statistic: $Z = \frac{\bar{x}-10}{2/\sqrt{25}} = \frac{0.92}{0.4} = 2.30$.

p-value: $p = P(Z' \geq 2.30) = 1 - \Phi(2.30) = 0.0107$.

At $\alpha = 0.05$: reject H_0. At $\alpha = 0.01$: do not reject. The p-value captures both decisions simultaneously.

Example 2: Multiple Testing with BH

A genomics study tests $m = 1000$ genes. Suppose 50 genes have true effects. Observed p-values include the ordered values $p_{(1)} = 0.00003$, $p_{(2)} = 0.00012$, ...

Apply BH at FDR level $\alpha = 0.05$. The BH threshold for the k-th ordered p-value is $k \cdot 0.05/1000 = k/20000$.

- $p_{(1)} = 0.00003 \leq 1/20000 = 0.00005$: yes.
- $p_{(2)} = 0.00012 \leq 2/20000 = 0.0001$: no.

So only the first gene is rejected at this threshold. In contrast, Bonferroni at $\alpha = 0.05$: threshold $= 0.05/1000 = 0.00005$, rejecting the same gene.

Now suppose $p_{(2)} = 0.00008$. Then $0.00008 \leq 0.0001$: yes. BH rejects both, while Bonferroni (threshold 0.00005) rejects only the first. This illustrates BH's greater power under many true alternatives.

Example 3: Discrete p-Value (Binomial)

Let $X \sim \text{Binomial}(10, p)$. Test $H_0 : p = 0.5$ vs $H_1 : p > 0.5$. Observed $X = 8$.

$$p\text{-value} = P_{0.5}(X \geq 8) = \sum_{k=8}^{10} \binom{10}{k} (0.5)^{10} = \frac{45 + 10 + 1}{1024} = \frac{56}{1024} \approx 0.0547.$$

At $\alpha = 0.05$: do not reject (barely). Note that the p-value is conservative (super-uniform) due to discreteness: $P_{0.5}(p(X) \leq u) \leq u$ with strict inequality for most u.

Cross-Links

- **AI:** p-values are used in A/B testing for model selection and feature evaluation. The multiple testing problem arises in neural architecture search, where many configurations are compared simultaneously.
- **Economics:** p-values are the standard reporting currency in empirical economics. The "p-hacking" debate (selective reporting of significant results) has driven reforms toward pre-registration and FDR control.
- **Linear Algebra:** For Gaussian test statistics, the p-value is a monotone function of the Mahalanobis distance from the null, connecting null distribution geometry to the probability scale.
- **Quantum:** In quantum state certification, a p-value can be defined for the null hypothesis that a prepared state equals a target state, using fidelity-based test statistics.

Structural Compression

Invariant: Under H_0, $p(X)$ is stochastically no smaller than $\text{Uniform}(0, 1)$; under a simple null with continuous T, it is exactly uniform.

Mechanism: The p-value is the survival function of T under P_0 evaluated at the observed $T(X)$; uniformity follows from the probability integral transform. The supremum over Θ_0 ensures validity for composite nulls.

Cross-System Echo: The p-value transforms a problem-specific test statistic into a universal probability scale, analogous to how quantile normalization maps arbitrary distributions to a reference distribution. In information theory, the p-value under the null is the "surprise" measured in probability units rather than bits.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with σ^2 known. Derive the p-value for the two-sided test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ and verify it is Uniform$(0, 1)$ under H_0.

2. Prove that if $p(X)$ is a valid p-value and $g : [0, 1] \to [0, 1]$ is non-decreasing with $g(u) \leq u$ for all u, then $g(p(X))$ is also a valid p-value. Give an example of such a g.

3. Compute the p-value for the LR test of $H_0 : \lambda = 1$ vs $H_1 : \lambda \neq 1$ given $X_1, \ldots, X_{20} \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$ with $\sum x_i = 28$, using both the exact distribution and the chi-squared approximation.

4. Demonstrate Lindley's paradox numerically: for $n = 100, 1000, 10000$, compute the sample mean $\bar{x}$ that gives a p-value of exactly 0.05, and compute the Bayes factor BF_{01} with prior $\mu \sim N(0, 1)$ under H_1.

5. Apply the Bonferroni and Benjamini–Hochberg procedures to the p-values $\{0.001, 0.013, 0.029, 0.032, 0.15, 0.45, 0.78\}$ at level $\alpha = 0.05$. Identify which hypotheses are rejected under each method.

6. Prove that for a discrete test statistic, the p-value is super-uniform: $\mathbb{P}_{\theta_0}(p(X) \leq u) \leq u$ with strict inequality for some u. Explain why randomized p-values can restore exact uniformity.

7. Argue, structurally, why the p-value is sufficient for performing a level-α test at any α, and explain how this relates to the p-value being a "minimal sufficient statistic" for the family of level-α rejection decisions indexed by $\alpha \in (0, 1]$.

Duality: Tests and Confidence Sets

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.6

This node formalizes the exact correspondence between hypothesis tests and confidence sets, unifying the two principal branches of frequentist inference. The duality is not merely an analogy: it is a bijection between families of level-α tests and level-$(1-\alpha)$ confidence sets. This result requires the testing framework (Node 5.1), estimation and confidence set theory (Nodes 4.1, 4.4), and draws on p-values (Node 5.5) for the continuous version of the correspondence. It serves as the capstone of Module 5, connecting testing to the broader inferential landscape.

Formal Definition

Setting

Let $X \sim P_\theta$, $\theta \in \Theta$. For each $\theta_0 \in \Theta$, suppose $\phi_{\theta_0}(X)$ is a level α test of

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0.$$

Test Inversion

The **test-inversion confidence set** is the set-valued function

$$C(X) = \big\{\theta_0 \in \Theta : \phi_{\theta_0}(X) = 0\big\},$$

the set of all parameter values not rejected by their respective level α tests.

Confidence Set Inversion

Given a confidence set $C(X)$ with level $1-\alpha$, define for each θ_0:

$$\phi_{\theta_0}(X) = \mathbf{1}\big(\theta_0 \notin C(X)\big).$$

Intuition

A confidence set collects all parameter values that are "compatible" with the observed data. A hypothesis test declares a specific parameter value "incompatible" when the data falls in the rejection region.

The duality states: a parameter value is in the confidence set if and only if the test at that value does not reject. This is not a coincidence of definitions — it is a structural identity. The coverage probability of the confidence set equals one minus the size of the test, because both quantities measure the same event: the true parameter is not falsely excluded.

The practical consequence is that any method for constructing tests immediately yields a method for constructing confidence sets, and vice versa. UMP tests yield shortest confidence sets; LR tests yield likelihood-based confidence regions; Wald tests yield Wald intervals.

Structural Role

Test–confidence duality unifies the two principal branches of frequentist inference. Confidence sets are not merely interval estimators; they are the geometric encoding of an entire family of hypothesis tests.

This duality explains why Wald, Score, and LR confidence intervals (Node 4.4) correspond exactly to inverting the Wald, Score, and LR test statistics: each test generates its confidence set by accumulating accepted parameter values.

The duality also connects to p-values (Node 5.5): the confidence set at level $1-\alpha$ consists of all θ_0 with p-value exceeding α.

Mathematical Development

Theorem – Test Inversion Produces a Confidence Set

Statement. If ϕ_{θ_0} is a level α test of $H_0 : \theta = \theta_0$ for each $\theta_0 \in \Theta$, then

$$C(X) = \{\theta_0 \in \Theta : \phi_{\theta_0}(X) = 0\}$$

satisfies

$$\mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha \quad \forall \theta \in \Theta.$$

Proof. Fix any $\theta \in \Theta$. Then

$$\mathbb{P}_\theta(\theta \in C(X)) = \mathbb{P}_\theta(\phi_\theta(X) = 0) = 1 - \mathbb{E}_\theta[\phi_\theta(X)].$$

Since ϕ_θ is a level α test of $H_0 : \theta = \theta$ (the true parameter), we have $\mathbb{E}_\theta[\phi_\theta(X)] \leq \alpha$. Therefore $\mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha$. $\square$

Theorem – Confidence Set Inversion Produces a Family of Tests

Statement. If $C(X)$ satisfies $\mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha$ for all θ, then for each θ_0,

$$\phi_{\theta_0}(X) = \mathbf{1}(\theta_0 \notin C(X))$$

is a level α test of $H_0 : \theta = \theta_0$.

Proof. For any $\theta_0 \in \Theta$:

$$\mathbb{E}_{\theta_0}[\phi_{\theta_0}(X)] = \mathbb{P}_{\theta_0}(\theta_0 \notin C(X)) = 1 - \mathbb{P}_{\theta_0}(\theta_0 \in C(X)) \leq 1 - (1 - \alpha) = \alpha. \quad \square$$

The Bijection

The two constructions are inverses. Starting from a family of tests $\{\phi_{\theta_0}\}$, forming $C(X)$, and then recovering tests via $\tilde{\phi}_{\theta_0}(X) = \mathbf{1}(\theta_0 \notin C(X))$ yields $\tilde{\phi}_{\theta_0} = 1 - (1 - \phi_{\theta_0}) = \phi_{\theta_0}$ for non-randomized tests.

For randomized tests, the correspondence generalizes: define $C_\alpha(X) = \{\theta_0 : \phi_{\theta_0}(X) < 1\}$ and the randomized inclusion event $\mathbb{P}(\theta_0 \in C(X)|X) = 1 - \phi_{\theta_0}(X)$.

Property Transfer Table

The duality transfers structural properties between tests and confidence sets:

Test Property	Confidence Set Property
Level α	Coverage $\geq 1-\alpha$
Exact level α	Exact coverage $1-\alpha$
Unbiased test	Unbiased confidence set
UMP test	Uniformly shortest confidence set
LR test	Likelihood-based confidence region
Equivariant test	Equivariant confidence set

UMP Tests Yield Shortest Confidence Sets

Proposition. If $\phi^*_{\theta_0}$ is UMP level α for each θ_0, then $C^*(X) = \{\theta_0 : \phi^*_{\theta_0}(X) = 0\}$ has the smallest expected Lebesgue measure among all level $(1-\alpha)$ confidence sets derived from level α tests.

Proof sketch. Let $C(X)$ be any other $(1-\alpha)$ confidence set derived from level α tests $\{\phi_{\theta_0}\}$. The expected length is $\mathbb{E}_\theta[\lambda(C(X))] = \int \mathbb{P}_\theta(\theta_0 \in C(X))\, d\theta_0 = \int (1 - \mathbb{E}_\theta[\phi_{\theta_0}(X)])\, d\theta_0$. Since $\phi^*_{\theta_0}$ is UMP, $\mathbb{E}_\theta[\phi^*_{\theta_0}(X)] \geq \mathbb{E}_\theta[\phi_{\theta_0}(X)]$ for $\theta_0 \neq \theta$, giving $\mathbb{E}_\theta[\lambda(C^*(X))] \leq \mathbb{E}_\theta[\lambda(C(X))]$. $\square$

p-Value Formulation

The duality admits a clean p-value formulation. Let $p_{\theta_0}(X)$ be the p-value for testing $H_0 : \theta = \theta_0$. Then

$$C_\alpha(X) = \{\theta_0 : p_{\theta_0}(X) > \alpha\}$$

is a $(1-\alpha)$ confidence set. The confidence set is the upper level set of the p-value function viewed as a function of θ_0.

Worked Examples

Example 1: Normal Mean Confidence Interval via Test Inversion

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, σ^2 known. For each μ_0, the level α two-sided test of $H_0 : \mu = \mu_0$ rejects when

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}.$$

The acceptance region (non-rejection) is $|\bar{X} - \mu_0| \leq z_{\alpha/2}\sigma/\sqrt{n}$, i.e., $\mu_0 \in [\bar{X} - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}]$.

Therefore

$$C(X) = \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right]$$

is the standard $(1-\alpha)$ confidence interval, recovered by inverting the z-test.

Example 2: LR Confidence Region for Two Parameters

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, both unknown. The LR test of $H_0 : (\mu, \sigma^2) = (\mu_0, \sigma_0^2)$ rejects when $-2 \log \Lambda_n > \chi^2_{2,\alpha}$.

The LR confidence region is

$$C(X) = \left\{ (\mu_0, \sigma_0^2) : -2 \log \Lambda_n(\mu_0, \sigma_0^2) \leq \chi^2_{2,\alpha} \right\}.$$

Explicitly:

$$-2 \log \Lambda_n = n \log \frac{\sigma_0^2}{\hat{\sigma}^2} + \frac{n(\bar{X} - \mu_0)^2 + n\hat{\sigma}^2}{\sigma_0^2} - n,$$

where $\hat{\sigma}^2 = n^{-1} \sum (X_i - \bar{X})^2$. The confidence region is the set of (μ_0, σ_0^2) for which this expression does not exceed $\chi^2_{2,\alpha}$.

Example 3: One-Sided Test to One-Sided Confidence Bound

For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$, the UMP test of $H_0 : \mu \leq \mu_0$ at level α rejects when $\bar{X} > \mu_0 + z_\alpha / \sqrt{n}$.

Inverting: the set of μ_0 not rejected is $\mu_0 \geq \bar{X} - z_\alpha / \sqrt{n}$. This yields the one-sided $(1 - \alpha)$ confidence bound

$$\mu \geq \bar{X} - z_\alpha / \sqrt{n}$$

or equivalently, the confidence set $C(X) = [\bar{X} - z_\alpha / \sqrt{n}, \infty)$.

Cross-Links

- **AI:** In conformal prediction, prediction sets are constructed by inverting a family of tests based on nonconformity scores, directly applying the test–confidence duality in a distribution-free setting.
- **Economics:** Confidence intervals for treatment effects in causal inference are obtained by inverting tests of sharp null hypotheses (Fisher's exact test inversion), a direct application of this duality.
- **Linear Algebra:** The confidence ellipsoid $\{(\mu_0) : (\bar{X} - \mu_0)^\top \Sigma^{-1} (\bar{X} - \mu_0) \leq c\}$ is the sublevel set of a quadratic form; its boundary is the acceptance boundary of the corresponding chi-squared test.
- **Quantum:** Quantum confidence regions for state tomography are constructed by inverting likelihood ratio tests on the density matrix, with the Hilbert–Schmidt metric replacing the Euclidean metric.

Structural Compression

Invariant: Level α size control in testing and $(1 - \alpha)$ coverage in confidence sets are the same structural constraint expressed in dual coordinates.

Mechanism: The acceptance region of the test at θ_0 becomes the event $\{\theta_0 \in C(X)\}$; inclusion probability equals one minus test size. The correspondence is a bijection for non-randomized tests and extends to randomized tests via probabilistic inclusion.

Cross-System Echo: In decision theory, this duality is the frequentist instance of the general duality between loss functions and risk sets. In convex optimization, the primal (confidence set as intersection of half-spaces) and dual (test as separation hyperplane) perspectives mirror the test–confidence correspondence. In Bayesian inference, credible sets are the posterior analogue: coverage is guaranteed by posterior mass rather than sampling frequency, but the structural role is identical.

Exercises

1. Derive the $(1-\alpha)$ confidence interval for the parameter p of a Bernoulli distribution by inverting the exact binomial test. This is the Clopper–Pearson interval.
2. Prove that if $C(X)$ is a $(1-\alpha)$ confidence set and $C'(X) \subseteq C(X)$ is a strict subset satisfying $\mathbb{P}_\theta(\theta \in C'(X)) \geq 1-\alpha$ for all θ, then the family of tests derived from $C'(X)$ has power no less than the family derived from $C(X)$.
3. For $X \sim N(\mu, 1)$, construct the confidence set by inverting the one-sided UMP test of $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$. Compare the result to the two-sided confidence interval.
4. Show that the Wald confidence interval $\hat{\theta} \pm z_{\alpha/2} \cdot \mathrm{se}(\hat{\theta})$ is obtained by inverting the Wald test $|\hat{\theta} - \theta_0|/\mathrm{se}(\hat{\theta}) > z_{\alpha/2}$.
5. Prove that inverting a family of unbiased level α tests yields an unbiased confidence set, i.e., $\mathbb{P}_\theta(\theta' \in C(X)) \leq 1-\alpha$ for all $\theta' \neq \theta$.
6. Consider the Score test for $H_0 : \theta = \theta_0$ in a one-parameter model. Write the confidence set obtained by inverting this test and compare it to the LR and Wald confidence sets in a specific example (e.g., Poisson mean).
7. Argue, structurally, why the test–confidence duality is an instance of a broader principle in mathematics: the duality between points and hyperplanes (or between membership and separation), and identify how this geometric perspective clarifies the relationship between coverage probability and rejection probability.