p-Values

Statistics · Module 5 — Hypothesis Testing

Node 5.5

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 $\text{Uniform}(0, 1)$ 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 $p(X) = 1 - U \sim \text{Uniform}(0,1)$. $\square$

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. $\square$

Connection to Level α Tests

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

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

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 $H_1 : \mu \ne 0$ 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 $H_{0,1}, \dots, H_{0,m}$ are tested simultaneously with p-values $p_1, \dots, p_m$, the probability of at least one false rejection (familywise error rate, FWER) can far exceed α.

Bonferroni correction. Reject $H_{0,i}$ if $p_i \le \alpha/m$. Then FWER $\le \alpha$ 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: $p_{(1)} \le \cdots \le p_{(m)}$. 2. Find the largest k such that $p_{(k)} \le k\alpha/m$. 3. Reject all $H_{0,(i)}$ for $i \le k$.

Theorem (Benjamini–Hochberg, 1995). If the p-values corresponding to the true nulls are independent, then the BH procedure controls FDR $\le \alpha \cdot m_0/m \le \alpha$, 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 $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ 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 $X_1, \dots, X_{20} \overset{\text{iid}}{\sim}$ Poisson(λ) 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 $\alpha \in (0, 1]$.

Statistics · Module 5 — Hypothesis Testing · Node 5.5
Concepts: hypothesis-testing
Prerequisites: 5.1, 5.2
Enables: 5.6

hbar.university — own.your.cognition