Hypothesis Testing

Andrew Nobel

September, 2020

Hypothesis Testing

Ingredients

- Family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ of densities on sample space $\mathcal{X}$
- Distinguished subset $\Theta_0 \subseteq \Theta$. Let $\Theta_1 = \Theta_0^c$

Goal: Use observation $X \sim f(x|\theta) \in \mathcal{P}$ to distinguish between hypotheses

- $\text{H}_0 : \theta \in \Theta_0$ Null hypothesis
- $\text{H}_1 : \theta \in \Theta_1$ Alternative hypothesis

Asymmetry: Null hypothesis represents status quo. Reject H_0 (accept H_1) only when there is significant evidence to do so

- (Medicine) H_0: New drug/procedure no better than existing drug/procedure
- (Quality control) H_0: Proportion of defectives acceptable
- (Criminal trial) H_0: Defendant is innocent

Some Terminology

Definition: H_0 is *simple* if $|\Theta_0| = 1$, *composite* otherwise. For $\Theta \subseteq \mathbb{R}$ null H_0 is

- one-sided if $\Theta_0 = \{\theta \in \Theta : \theta \leq \theta_a\}$ or $\Theta_0 = \{\theta \in \Theta : \theta \geq \theta_b\}$
- two-sided if $\Theta_0 = \{\theta \in \Theta : \theta \leq \theta_a \text{ or } \theta \geq \theta_b\}$

Example: $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\theta, 1)$ with $\theta \in \Theta \subseteq \mathbb{R}$

(a) $\Theta = \{0, 1\}$. Test $\mathrm{H}_0 : \theta = 0$ vs. $\mathrm{H}_1 : \theta = 1$

(b) $\Theta = [0, \infty)$. Test $\mathrm{H}_0 : \theta = 0$ vs. $\mathrm{H}_1 : \theta > 0$

(c) $\Theta = \mathbb{R}$ with θ_0 fixed. Possible tests

$$\mathrm{H}_0 : \theta \leq \theta_0 \text{ vs. } \mathrm{H}_1 : \theta > \theta_0$$

$$\mathrm{H}_0 : \theta = \theta_0 \text{ vs. } \mathrm{H}_1 : \theta \neq \theta_0$$

Hypothesis Tests

Hypothesis test: Specified by a pair (T, R)

- Test statistic $T : \mathcal{X} \to \mathcal{T}$
- Rejection region $R \subseteq \mathcal{T}$
- Decision rule $d(x) = \mathbb{I}(T(x) \in R)$

Test rejects H_0 if $T(x) \in R$, accepts H_0 if $T(x) \in R^c$

Note

- Test partitions $\mathcal{X}$ into $\mathcal{X}_1 = \{x : T(x) \in R\}$ and $\mathcal{X}_0 = \{x : T(x) \notin R\}$
- Thus we can always take $T(X) = X$ and $R \subseteq \mathcal{X}$
- Rejection region R will depend on H_0, e.g., one-sided vs. two-sided

Likelihood Ratio Tests

Idea: Compare maximum of $L(\theta|x)$ over null Θ_0 and over full parameter space Θ

Definition: The *likelihood ratio* test statistic is

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta|x)}{\sup_{\theta \in \Theta} L(\theta|x)} \in [0, 1]$$

The *likelihood ratio test* (LRT) rejects H_0 if $\lambda(x) \leq c$ for selected $c \in [0, 1]$

Note

- Large values of $\lambda(x)$ favor H_0, small values favor H_1
- $\lambda(x) = 1$ iff $\sup_{\theta \in \Theta_0} L(\theta|x) \geq \sup_{\theta \in \Theta_1} L(\theta|x)$
- $\lambda(x) = L(\hat{\theta}_0|x)/L(\hat{\theta}|x)$ where $\hat{\theta}_0 =$ MLE for Θ_0 and $\hat{\theta} =$ MLE for Θ

LRT for Normal Mean, Known Variance

Let $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\theta, \sigma^2)$ with $\theta \in \mathbb{R}$ and σ^2 known. Consider testing

$$\text{H}_0 : \theta = \theta_0 \;\; \text{vs.} \;\; \text{H}_1 : \theta \neq \theta_0$$

Fact. LRT $(\lambda, [0, c])$ is equivalent to a two-sided z-test, that is, $\lambda(x) \leq c$ iff

$$T(x) \;=\; \frac{|\overline{x} - \theta_0|}{\sigma/\sqrt{n}} \;\geq\; z(c)$$

for some number $z(c)$

LRT for Normal Mean, Unknown Variance

Given $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ unknown. Consider testing

$$\mathsf{H}_0 : \mu \leq \mu_0 \ \text{ vs. } \ \mathsf{H}_1 : \mu > \mu_0$$

Terminology: Variance σ^2 is said to be a *nuisance parameter*. It is unknown but does *not* figure in the hypotheses being tested.

Fact. LRT $(\lambda, [0, c])$ is equivalent to a one-sided t-test, that is, $\lambda(x) \leq c$ iff

$$T(x) \ = \ \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \ \geq \ t(c)$$

for some number $t(c) \geq 0$

LRT and Sufficient Statistics

Let $T : \mathcal{X} \to \mathcal{T}$ be sufficient for θ with $T(X) \sim g(t|\theta)$ when $X \sim f(x|\theta)$. Define

- $\tilde{L}(\theta|t) = g(\theta|t)$ likelihood of θ given $T(X) = t$
- $\tilde{\lambda}(t) =$ Likelihood ratio test statistic for T based on $\tilde{L}(\theta|t)$

Fact: For each $x \in \mathcal{X}$ we have $\tilde{\lambda}(T(x)) = \lambda(x)$.

Proof: Follows from factorization theorem

Upshot: We can always construct a LRT based on a sufficient statistic

Evaluating Test Performance

Test $\mathrm{H}_0 : \theta \in \Theta_0$ vs. $\mathrm{H}_1 : \theta \in \Theta_1$ involves binary decision rule $d : \mathcal{X} \to \{0, 1\}$

- $d(x) = 0$ means we accept H_0 (equivalently, reject H_1)
- $d(x) = 1$ means we reject H_0 (equivalently, accept H_1)

Under the 0/1 loss a rule d can make two kinds of errors (each with loss 1)

- Type I: Reject H_0 when H_0 true (i.e. $d(x) = 1$ when $\theta \in \Theta_0$)
- Type II: Accept H_0 when H_1 true (i.e. $d(x) = 0$ when $\theta \in \Theta_1$)

Truth / Decision	H_0	H_1
H_0	Correct	Type I
H_1	Type II	Correct

Power Function of a Test

Definition: The *power function* $\beta : \Theta \to [0, 1]$ of a test (X, R) or rule $d : \mathcal{X} \to \{0, 1\}$ is

$$\beta(\theta) \;=\; \mathbb{P}_\theta(X \in R) \;=\; \mathbb{P}_\theta(d(X) = 1)$$

Power function contains all relevant information about error probabilities of the test

- If $\theta \in \Theta_0$ then $\beta(\theta) = P_\theta$(Type I error)
- If $\theta \in \Theta_1$ then $\beta(\theta) = 1 - P_\theta$(Type II error), *power* of test at alternative θ

Competing Goals

- $\beta(\theta)$ close to zero for $\theta \in \Theta_0$ (small Type I error)
- $\beta(\theta)$ close to one for $\theta \in \Theta_1$ (large power against alternatives)

Level, Size, Unbiasedness

Standard approach: Maximize power of test subject to bound on Type I error

Definition: Let (X, R) be a test having power function $\beta(\theta)$ and let $\alpha \in [0, 1]$

- (X, R) is a *level*-α test if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$
- (X, R) is a *size*-α test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$
- (X, R) is *unbiased* if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \inf_{\theta \in \Theta_1} \beta(\theta)$

Examples

Ex 1. (Binomial mean) Let $X \sim \text{Bin}(n, \theta)$ with $\theta \in [0, 1]$. Consider testing

$$\text{H}_0 : \theta \leq 1/2 \ \text{ vs. } \ \text{H}_1 : \theta > 1/2$$

Smaller values of X favor H_0 so consider rejection regions $R_k = \{k, k+1, \ldots, n\}$.

Of interest: Power function $\beta_k(\theta)$ of test (X, R_k).

Ex 2. (Normal mean) Let $X_1, \ldots, X_n$ be i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. Consider

$$\text{H}_0 : \mu \leq \mu_0 \ \text{ vs. } \ \text{H}_1 : \mu > \mu_0$$

In this case LRT is equivalent to (T, R_c) where

$$T(x) \ = \ \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \qquad\qquad R_c = [c, \infty)$$

Of interest: Power function $\beta_c(\theta)$ of test (T, R_c).

Uniformly Most Powerful Tests

Definition: Test (X, R) is a uniformly most powerful (UMP) level α test for H_0 vs. H_1 if

(1) (X, R) has level α

(2) For any other level α test (X, R')

$$\beta(\theta) = \mathbb{P}_\theta(X \in R) \geq \mathbb{P}_\theta(X \in R') = \beta'(\theta) \quad \text{for all } \theta \in \Theta_1$$

In other words: (X, R) has maximum power among all level α tests.

Neyman-Pearson Theorem

Setting: Family $\mathcal{P} = \{f_0, f_1\}$ with two densities on $\mathbb{R}$, parameters $\Theta = \{0, 1\}$

- Given $X \sim f_\theta$ wish to test

$$\text{H}_0 : \theta = 0 \;\text{ vs. }\; \text{H}_1 : \theta = 1$$

- Natural test statistic is the likelihood ratio

$$T(x) \;=\; \frac{f_1(x)}{f_0(x)} \;=\; \frac{\text{likelihood of } \theta = 1 \text{ given } x}{\text{likelihood of } \theta = 0 \text{ given } x}$$

Thm: Let $\mathcal{P} = \{f_0, f_1\}$ and let (X, R) be a size α test for $\mathrm{H}_0 : \theta = 0$ vs. $\mathrm{H}_1 : \theta = 1$ that is based on thresholding f_1/f_0. More precisely, for some $\tau > 0$

$$\mathbb{P}_0(X \in R) = \alpha \;\text{ and }\; \{f_1 > \tau f_0\} \subseteq R \subseteq \{f_1 \geq \tau f_0\}$$

Sufficiency: (X, R) is a UMP level α test

Necessity: If (X, R') is any UMP level α test then (i) $\mathbb{P}_0(X \in R') = \alpha$ and (ii) R, R' differ only on the set where $f_1 = \tau f_0$

Note

- (X, R) rejects H_0 if $f_1(x)/f_0(x) > \tau$, accepts H_0 if $f_1(x)/f_0(x) < \tau$
- Value of τ and set $R \cap \{f_1 = \tau f_0\}$ depend on α.
- Conclusion (i) of necessity says that (X, R') is of size α
- Conclusion (ii) of necessity is $P_\theta((R \triangle R') \cap \{f_1 \neq \tau f_0\}) = 0$ for $\theta = 0, 1$

P-Values: General Definition

Definition: A valid p-value for a test of $\mathrm{H}_0 : \theta \in \Theta_0$ vs. $\mathrm{H}_1 : \theta \in \Theta_1$ is a function $p : \mathcal{X} \to [0, 1]$ such that for every $\alpha \in [0, 1]$

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(p(X) \leq \alpha) \;\leq\; \alpha$$

- By definition, $p(X)$ is stochastically larger than a $\mathrm{U}(0, 1)$ random variable whenever $X \sim P_\theta$ with $\theta \in \Theta_0$
- Interpret value of $p(x)$ as quantifying evidence against H_0 given observation $X = x$. Smaller values of $p(x)$ mean more evidence against H_0
- Definition ensures that test $(p(\cdot), [0, \alpha])$ has level α

Standard Definition

Setting: Testing $\mathsf{H}_0 : \theta \in \Theta_0$ vs. $\mathsf{H}_1 : \theta \in \Theta_1$ using statistic $S : \mathcal{X} \to \mathbb{R}$ with larger values favoring the alternative

- Observation $X \sim P_\theta$ yields data $x \in \mathcal{X}$, value $S(x)$ of test statistic
- Task: quantify the strength of evidence against H_0 provided by the observed value of the test statistic
- Consider probability that test statistic $S(X)$ exceeds observed value $S(x)$ under the null H_0. Define p-value

$$p(x) \;=\; \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(S(X) \geq S(x))$$

Fact: If $X \in \mathbb{R}$ has CDF F then $F(X) \stackrel{d}{\geq} \mathsf{U}[0,1]$

Proposition: The function $p(x)$ is a valid p-value

Examples

Ex 1. Observe $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 unknown. Wish to test

$$\mathsf{H}_0 : \mu = \mu_0 \text{ vs. } \mathsf{H}_1 : \mu \neq \mu_0$$

Ex 2. Observe $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 unknown. Wish to test

$$\mathsf{H}_0 : \mu \leq \mu_0 \text{ vs. } \mathsf{H}_1 : \mu > \mu_0$$

Bayesian Hypothesis Testing

Setting: Family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ of densities on $\mathcal{X}$, prior $\pi(\theta)$ on Θ. Consider

$$\mathsf{H}_0 : \theta \in \Theta_0 \;\text{ vs. }\; \mathsf{H}_1 : \theta \in \Theta_1$$

In the Bayesian setting, the relevant quantities are "conceptually straightforward"

- Prior probability of hypotheses H_0 and H_1

$$P(\mathsf{H}_0) = \int_{\Theta_0} \pi(\theta)\, d\theta \qquad P(\mathsf{H}_1) = \int_{\Theta_1} \pi(\theta) d\theta$$

- Posterior probability of hypotheses H_0 and H_1 *given data* $X = x$

$$P(\mathsf{H}_0 |\, x) = \int_{\Theta_0} \pi(\theta|x)\, d\theta \qquad P(\mathsf{H}_1 |\, x) = \int_{\Theta_1} \pi(\theta|x)\, d\theta$$

Choose between H_0 and H_1 based on posterior probabilities $P(\mathsf{H}_0 |\, x)$ and $P(\mathsf{H}_1 |\, x)$

Odds Ratios and the Bayes Factor

Definition

- Prior odds ratio $P(\mathsf{H}_1)/P(\mathsf{H}_0)$
- Posterior odds ratio $P(\mathsf{H}_1 \,|\, x)/P(\mathsf{H}_0 \,|\, x)$
- Bayes factor

$$B = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}} = \frac{P(\mathsf{H}_1 \,|\, x)\, P(\mathsf{H}_0)}{P(\mathsf{H}_0 \,|\, x)\, P(\mathsf{H}_1)}$$

Interpretation: Bayes factor measures strength of evidence for H_1 over H_0

- posterior odds $=$ Bayes factor $\times$ prior odds
- $B =$ posterior odds ratio if $P(\mathsf{H}_0) = P(\mathsf{H}_1)$
- Bayes factor can be written as *weighted* likelihood ratio $B = P(x \,|\mathsf{H}_1)/P(x \,|\mathsf{H}_0)$