p-Values, Confidence Regions

Outline

1) p-Values
2) Confidence regions
3) (Mis-)interpreting tests

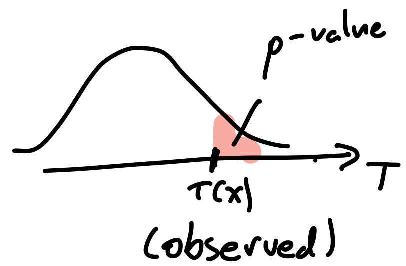
p-Values

<u>Informal definition:</u> Suppose $\phi(X)$ rejects for large values of $T(X)$.

$p(x)$ = "Null probability that $T(X)$ is as large or larger than what we observed"

$$= \text{``}\mathbb{P}_{H_0}\left(T(X) \geq T(x)\right)\text{''}$$

$$= \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\left(T(X) \geq T(x)\right)$$



<u>Ex</u> $X \sim \text{Binom}(n, \theta)$ $\quad H_0: \theta \leq 0.5$ vs $H_1: \theta > 0.5$

One-sided test rejects for large X

$$p(x) = \mathbb{P}_{0.5}(X \geq x) = \sup_{\theta \leq 0.5} \mathbb{P}_\theta(X \geq x)$$

<u>Ex</u> $X \sim N(\theta, 1)$ $\quad H_0: \theta = 0$ vs. $H_1: \theta \neq 0$

Two-sided test rejects for large $T(X) = |X|$

$$\left(\Leftrightarrow \phi_\alpha(X) = 1\{ |X| > z_{\alpha/2} \} \right)$$

The two-sided p-value is $p(X)$ where

$$p(x) = \mathbb{P}_0(|X| > |x|)$$

$$= 2(1 - \Phi(|x|))$$

Formal definition: $\mathcal{P}$, Θ_0, Θ.

Assume we have a test ϕ_α for each significance level, $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \phi_\alpha(X) \le \alpha$

(non-randomized case: $\phi_\alpha = 1\{x \in R_\alpha\}$)

Assume tests are monotone in α:
if $\alpha_1 \le \alpha_2$ then $\phi_{\alpha_1}(x) \le \phi_{\alpha_2}(x)$

(non-randomized: $R_{\alpha_1} \subseteq R_{\alpha_2}$)

Then $p(x) = \sup\{\alpha : \phi_\alpha(x) < 1\}$

$(= \sup\{\alpha : x \notin R_\alpha\})$

(possible to define randomized p-value but not worth it)

For $\theta \in \Theta_0$, $\mathbb{P}_\theta(p(X) \le \alpha) = \mathbb{P}(\sup\{\tilde{\alpha} : \phi_{\tilde{\alpha}}(X) < 1\} \le \alpha)$

$$\le \inf_{\tilde{\alpha} > \alpha} \underbrace{\mathbb{P}_\theta(\phi_{\tilde{\alpha}}(X) = 1)}_{\le \tilde{\alpha}} \le \alpha$$

$\Rightarrow$ p-value stochastically dominates $U[0,1]$

If ϕ_α rejects for large $T(X)$,
reduces to original definition.

Note the p-value is defined relative to

- the model & null hyp.,
- the data, AND
- the choice of test

Ex $X \sim \mathrm{Exp}(\theta)$ $\quad H_0: \theta = 1$ vs $H_1: \theta \neq 1$

We can use equal-tailed test

or UMPU test

For $X > 1$:

Equal-tailed: $p(x) = 2 \cdot \mathbb{P}_1(X \geq x) = 2e^{-x}$

UMPU: $p(x) = \alpha$ for which $c_2(\alpha) = x$

Ex $X \sim N_d(\theta, I_d)$ $\quad H_0: \theta = 0$ vs $H_1: \theta \neq 0$

We can use $T_1(X) = \|X\|^2$ (χ^2 test)

or $T_2(X) = \|X\|_\infty$ (max test)

$= \max_i |X_i|$

Very different p-values / power if d large

(choice reflects belief about whether θ is sparse)

Confidence Sets

[Accept/reject decision only so interesting:
- usually we care how big θ is
- tiny p-value doesn't imply big θ
  (big p-value doesn't imply small θ either)]

<u>Def</u> $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$

$C(X)$ is a <u>$1-\alpha$ confidence set</u> for $g(\theta)$ if

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

subject

verb

object

We say $C(X)$ <u>covers</u> $g(\theta)$ if $C(X) \ni g(\theta)$

$\mathbb{P}_\theta(C(X) \ni g(\theta))$ is <u>coverage probability</u>

$\inf_\theta \mathbb{P}_\theta(C \ni g(\theta))$ is <u>conf. level</u>

<u>Notes</u>
- $C(X)$ is random, not $g(\theta)$
- Often misinterpreted as Bayesian guarantee
- Say "$C(X)$ has a 95% chance of covering"
  NOT "$g(\theta)$ has a 95% chance of being in C"
  NEVER "95% chance $g(\theta) \in [0.5, 1.5]$" (e.g.)

Duality of Testing & Confidence Sets

Suppose we have a level-α test $\phi(x;a)$ of $H_0: g(\theta) = a$ vs. $H_1: g(\theta) \neq a$, $\forall a \in g(\Theta)$

We can use it to make a confidence set for $g(\theta)$:

Let $C(X) = \{a : \phi(x;a) < 1\}$

$=$ "all non-rejected values of θ"

Then $\mathbb{P}_\theta\left(C(X) \not\ni g(\theta)\right) = \mathbb{P}_\theta\left(\phi(x; g(\theta)) = 1 \right)$

$\leq \alpha \qquad \forall \theta$

Alternatively, suppose $C(X)$ is a $1-\alpha$ confidence set for $g(\theta)$.

We can use C to contruct a test $\phi(X)$ of $H_0: g(\theta) = a$ vs. $H_1: g(\theta) \neq a$

$$\phi(X) = 1\{a \notin C(X)\}$$

For θ s.t. $g(\theta) = a$:

$$\mathbb{E}_\theta\, \phi(X) = \mathbb{P}_\theta\left(C(X) \not\ni g(\theta)\right) \leq \alpha$$

This is called <u>inverting the test</u>

Confidence interval for median

Nonparametric model $X_1, \dots, X_n \overset{iid}{\sim} F$, F any cdf

$$g(F) = \text{median}(F) = F^{-1}(1/2) \quad \text{(assume well-defined)}$$

Two-sided sign test:

$$H_0: g(F) = \mu \iff F(\mu) = 1/2$$

$$\text{vs} \quad H_1: g(F) \neq \mu \iff F(\mu) \neq 1/2$$

$$S(X;\mu) = \#\{X_i > \mu\} \sim \text{Binom}(n, \underbrace{1 - F(\mu)}_{= 1/2 \text{ iff } H_0 \text{ true}})$$

Reject for $T(X;\mu) = |S(X;\mu) - n/2| > c_\alpha$ ← no dependence on μ

e.g. $n = 100$ $c_\alpha = 5$, reject if $S(X) > 55$

$$\mu \in C(X) \iff |S(X;\mu) - n/2| \le c_\alpha$$

$$\iff \#\{X_i > \mu\} \in [n/2 - c_\alpha, n/2 + c_\alpha]$$

$$\iff \mu \in [X_{(n/2 - c_\alpha)}, X_{(n/2 + c_\alpha)}]$$

Confidence Intervals / Bounds

If $C(X) = [C_1(X), C_2(X)]$ we say $C(X)$ is a confidence interval (CI)

$C(X) = [C_1(X), \infty)$: lower conf. bd. (LCB)

$C(X) = (-\infty, C_2(X)]$: upper conf bd. (UCB)

We usually get LCB / UCB by inverting a one-sided test in appropriate direction

Called uniformly most accurate (UMA) if test UMP

Get CI by inverting a two-sided test

Called UMAU if test is UMPU

Ex $\quad X \sim \text{Exp}(\theta) = \frac{1}{\theta} e^{-x/\theta} \qquad x > 0, \ \theta > 0$

CDF $\quad \mathbb{P}_\theta(X \le x) = 1 - e^{-x/\theta}$

<u>LCB</u>: Invert test for $H_0: \theta \le \theta_0$

Solve $\quad \alpha = \mathbb{P}_{\theta_0}(X > c(\theta_0)) = e^{-c(\theta_0)/\theta_0}$

$$c(\theta_0) = \theta_0 \log(1/\alpha) \quad (>0)$$

$$X \le c(\theta_0) \Rightarrow \theta_0 \ge \frac{X}{-\log \alpha}$$

$$C(X) = \left[\frac{X}{-\log\alpha}, \infty\right)$$

<u>UCB</u>: Similar, $\quad C(X) = \left(-\infty, \frac{X}{-\log(1-\alpha)}\right]$

<u>Equal-tailed CI</u>:

Invert equal-tailed test of $H_0: \theta = \theta_0$

$$\underbrace{\phi_\alpha^{ET}(X)}_{\substack{\text{equal-tailed} \\ H_0: \theta = \theta_0}} = \underbrace{\phi_{\alpha/2}^{\ge\theta_0}(X)}_{H_0: \theta \ge \theta_0} + \underbrace{\phi_{\alpha/2}^{\le\theta_0}(X)}_{H_0: \theta \le \theta_0}$$

$$\Rightarrow C(X) = \left[\frac{X}{-\log\alpha/2}, \infty\right) \cap \left(-\infty, \frac{X}{-\log(1-\alpha/2)}\right]$$

$$= \left[\frac{X}{-\log\alpha/2}, \frac{X}{-\log(1-\alpha/2)}\right]$$

Similar for UMPU 2-sided test

(Mis-)Interpreting Hypothesis Tests

Hypothesis tests ubiquitous in science

Common misinterpretations:

1) $p < 0.05$ therefore "there is an effect"
   or "the effect size = the estimate"

2) $p > 0.05$ therefore "there is no effect"

3) $p = 10^{-6}$ therefore "the effect is huge"

4) $p = 10^{-6}$ therefore "the data are signif."
   and everything about our model is correct in most naive interp.

5) Effect CI for men is $[0.2, 3.2]$,
   for women is $[-0.2, 2.8]$ therefore
   "there is an effect for men and not for women."

Dichotomous test doesn't eliminate uncertainty
(CIs usually less misleading to novices)

How to interpret testing

Learning about the world from data is not easy or automatic!

Hypothesis tests let us ask specific questions about specific data sets under specific modeling assumptions, using specific testing method.

All of these choices bear on the interpretation.

Top-tier medical journals let people publish claims, reporting p-values without saying what model was used or what test was employed

Pretty bad when you think about it!

Hyp. tests can be a good companion to critical thinking, never a substitute

"All models are wrong, some are useful" but need experience and theory to understand when assumptions do or don't cause real trouble

Conceptual Objections

Q1: Why should I test $H_0: \theta = 0$? No θ is ever exactly 0.

A1: a) Test $H_0: |\theta| \leq \delta$ if you want

If $\text{se.}(\hat{\theta}) \gg \delta$, not much difference.

b) Most two-sided tests justify <u>directional inference</u>:

"If $T > c_\alpha$ declare $\theta > 0$, if $T < c_1$ declare $\theta < 0$" with $\mathbb{P}(\text{false claim}) \leq \alpha$

c) Harder to answer in non-parametric problems, e.g. $H_0: P = Q$ vs $H_1: P \neq Q$ for perm. test, but alternative frameworks like Bayes force very strong assumptions on us.

Q2: People only like frequentist results like p-values, CIs because they mistake them for Bayesian results.

95% chance $C(X) \ni \theta$ is misinterpreted as a claim about $p(\theta \mid X)$.

A2: True, but subjective Bayesian results often misinterpreted as "<u>the</u> posterior dist. of θ" when reall should be "<u>my</u> posterior opinion about θ"