Statistics 210A Lecture 16 Notes

Daniel Raban

October 19, 2021

1 Confidence Sets and Philosophy of Hypothesis Testing

1.1 Recap: hypothesis tests and p-values

We have been studying hypothesis testing, taking a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and distinguishing between two submodels $H_0 : \theta = \Theta_0$ and $H_1 : \theta = \Theta_1$. The hypothesis test is defined by its **critical function** $\phi(x) \in [0, 1]$.

In a simple null vs simple alternative hypothesis, we saw that it was optimal to reject for large $\frac{p_1}{p_0}(X)$. When we have one real parameter ($\Theta = R$, $\Theta = (0, \infty)$, etc.), this let us analyze 1-sided tests $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$. If $\frac{p_2}{p_1}$ is increasing in $T(x)$, for all $\theta_2 > \theta_1$ (MLR), then the UMP test rejects rejects for large $T(X)$. This is also valid if $T(X)$ is stochastically increasing in θ.

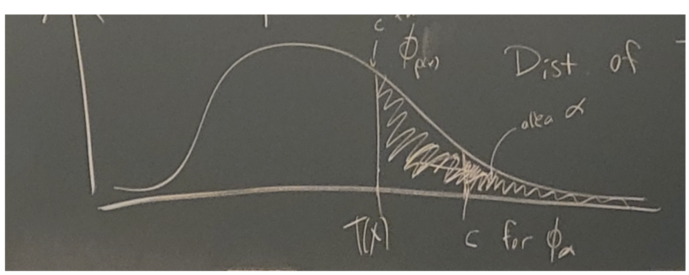
For 2-sided tests, i.e. $H_0 : \theta = \theta_0$ vs $H_1 \theta \neq \theta$ (or $H_0 : \theta_1 \leq \theta \leq \theta_2$ vs $H_1 : \theta < \theta_1$ or $\theta > \theta_2$), a 2-sided test rejects for extreme $T(X)$, where $T(x)$ is some test statistic. Here are two ways of making a two tailed test:

- Equal-tailed: Require $\mathbb{P}_{\theta_0}(T(X) > c_2) = \mathbb{P}_{\theta_0}(T(X) < c_1) = \alpha/2$.
- Unbiased: Require $\mathbb{P}_{\theta_0}(T(X) < c_1 \text{ or } > c_2) = \alpha$.

Example 1.1. For an exponential family, the 2-tailed unbiased test is UMPU.

The **p-value** is the level of α for which the test barely rejects:

$$\begin{aligned} p(x) &= \min\{\alpha : \phi_\alpha(x) = 1\} \\ &\overset{\text{often}}{=} \mathbb{P}_{\theta_0}(T(X) \geq T(x)). \end{aligned}$$



The p-value is defined with respect to a family of tests.

For $\theta \in \Theta_0$,

$$\mathbb{P}_\theta(p(X) \leq \alpha) = \mathbb{P}_\theta(\phi_\alpha(X) = 1) \leq \alpha,$$

so $p(X)$ stochastically dominates the uniform distribution on $(0, 1)$.

1.2 Confidence sets

Often, the effect size is a much more relevant question of whether there is an effect or in what direction the effect is.

Definition 1.1. In a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with an estimand $g(\theta)$, $C(X)$ is a $1 - \alpha$ **confidence set** for $g(\theta)$ if

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

In other words, the probability that we picked a set containing the estimand is $\geq 1 - \alpha$.

Remark 1.1. Note that we have written $C(X) \ni g(\theta)$, rather than the mathematically equivalent $g(\theta) \in C(X)$. This is because $g(\theta)$ is fixed; it is just the bullseye we are shooting for. $C(X)$ is the randomly determined object. People misinterpret this as a statement about $g(\theta)$ conditional on the data, which does not make sense from a frequentist viewpoint.

This should not be called a "confidence" set because confidence is a Bayesian notion. This should really be called an "interval estimate" instead.

1.3 Duality of confidence sets and testing

How do we make confidence sets? Suppose for every value a, we have a level-α test $\phi(x; \alpha)$ for $H_0 : g(\theta) = a$ vs $H_1 : g(\theta) \neq a$. Let

$$\begin{aligned} C(X) &= \{a : \phi(X; a) < 1\} \\ &= \{\text{all non-rejected values}\}. \end{aligned}$$

Then for every θ,

$$\mathbb{P}_\theta(C(X) \not\ni g(\theta)) = \mathbb{P}_\theta(\phi(X; a) = 1) \leq \alpha.$$

Note that the two appearances of θ on the left hand side need to be the same θ.

Remark 1.2. Why don't we need a correction for multiple testing, if we are making uncountably many tests? There is only one true null, so we only have 1 chance to make a type I error.

The above procedure is called **inverting a test** to get a confidence set. We can go the other way: We could rekect $H_0 : \theta \in \Theta_0$ if $C(X) \cap \Theta_0 = \varnothing$. For $\theta \in \Theta_0$,

$$\mathbb{P}_\theta(\text{test rejects}) = \mathbb{P}_\theta(\theta \notin C(X)) \leq \alpha.$$

Example 1.2. A **confidence interval** is a confidence set $C(X)$ which is an interval $[C_1(X), C_2(X)]$. This is usually obtained by inverting a two-sided test.

Example 1.3. An **upper confidence bound** is $C_2(X)$, where $C(X) = (-\infty, C_2(X)]$, and a **lower confidence bound** is $C_1(X)$, where $C(X) = [X_1(X), \infty)$. These are usually obtained by inverting a one-sided test.

Definition 1.2. A upper/lower confidence bound is called **uniformly most accurate (UMA)** if it inverts a UMP test. A confidence interval is called **UMAU** if it inversets a UMPU test.

Example 1.4. Suppose we observe $X \sim \text{Exp}(\theta) = \frac{1}{\theta} e^{-x/\theta}$ with $\theta > 0$. The CDF is $\mathbb{P}_\theta(X \leq x) = 1 - e^{-x/\theta}$.

- To get a lower confidence bound for θ, invert the one-sided test for $H_0 : \theta \leq \theta_0$. Solve

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

  to get

  $$c(\theta_0) = \theta_0(-\log \alpha) > 0.$$

  Now

  $$\begin{aligned} \phi(x; \theta_0) = 0 &\iff X \leq c(\theta_0) \\ &\iff \theta_0 \geq \frac{X}{-\log \alpha}. \end{aligned}$$

  So the confidence region is $C(X) = [\frac{X}{-\log \alpha}, \infty)$.

- For an upper confidence bound, a similar argument gives $C(X) = (-\infty, \frac{X}{-\log(1-\alpha)}]$.

- For a confidence interval derived from inverting an equal-tailed test, the equal-tailed test is

  $$\phi^{2T}\alpha(X; \theta_0) = \phi_{\alpha/2}^{\geq \theta_0}(X; \theta_0) + \phi_{\alpha/2}^{\leq \theta_0}(X; \theta_0),$$

  where these tests test $H_0 : \theta = \theta_0$, $H_0 : \theta \geq \theta_0$, and $H_0 : \theta \leq \theta_0$, respectively. Then the confidence interval is

  $$\begin{aligned} C(X) &= \big[\tfrac{X}{-\log(\alpha/2)}, \infty\big) \cap \big(-\infty, \tfrac{X}{-\log(1-\alpha/2)}\big] \\ &= \big[\tfrac{X}{-\log(\alpha/2)}, \tfrac{X}{-\log(1-\alpha/2)}\big]. \end{aligned}$$

1.4 Philosophy: misinterpreting hypothesis tests and objections to hypothesis testing

Here are some ways people misinterpret hypothesis tests:

1. If $p < 0.05$, then "there is an effect."
2. If $p > 0.05$, then "there is no effect."

The hypothesis test does not eliminate uncertainty; it just describes or quantifies the uncertainty.

3. If $p = 10^{-6}$, then " the effect is huge."
4. If $p = 10^{-6}$, then "the data are significant," and everything about our model is incorrect.
5. The effect confidence interval for men is $[0.2, 3.2]$ and for women is $[-0.2, 3.8]$, therefore "there is an effect for men and not for women."

Hypothesis tests ask specific questions about specific data sets under specific modeling assumptions using a specific testing method. Top tier medical journals, for example, let people publish claims by reporting p-values without saying what their model was or how they tested the data. But even if we do hypothesis testing right, here are some more objections:

1. Why should we ever test $H_0 : \theta = 0$? Maybe exact zero effects don't exist! Here are some responses:
   (a) One answer is that we could test something else, for example $H_0 : |\theta| \leq \delta$, where δ is some minimum effect size we care about. However, in a $N(\theta, \sigma^2)$ model, the power of this δ test $= \alpha + O((\delta/\sigma)^2)$
   (b) Usually, directional claims are justified.
   (c) In a 2-sample problem, we can test $H_0 : P = Q$ vs $H_1 : P \neq Q$, so this is harder to answer in non-parametric problems.
2. People only like frequentist results like p-values and confidence intervals because they mistake them for Bayesian results.
3. p-values ignore $\mathbb{P}(\text{Data} \mid H_0)$ and only look at $\mathbb{P}(\text{Data} \mid H_1)$. The data could be more likely under the null than under the alternative.
4. Maybe we should use something else instead of hypothesis testing, since scientists often misuse hypothesis tests.