Lecture 21. Hypothesis Testing II

December 7, 2011

In the previous lecture, we defined a few key concepts of hypothesis testing and introduced the framework for parametric hypothesis testing. In the parametric hypothesis testing framework, we assume that the pdf/pmf of a random variable/vector of interest, X, is known up to a finite dimensional parameter. Denote the pdf/pmf by $f_X(x, \theta)$ where θ is the parameter and θ is assumed to live in the parameter space $\Theta \subseteq R^{d_x}$. The hypotheses restrict the parameter to subsets of Θ:

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1 := \Theta / \Theta_1. \tag{1}$$

Assume that we have a random n-sample, $\mathbb{X} = \{X_1, ..., X_n\}$. A test is a rejection region, C_R, which typically take the form:

$$C_R = \{\mathbb{X} : T(\mathbb{X}) > c\}, \tag{2}$$

where $T(\mathbb{X})$ is a test statistic and c is a critical value. We defined the power function of the test to be:

$$\gamma(\theta) := \Pr{}_\theta(T(\mathbb{X}) > c) \quad \theta \in \Theta. \tag{3}$$

The size of the test is the maximum null rejection probability:

$$Sz = \max_{\theta \in \Theta_0} \gamma(\theta).$$

A test is said to be of significance level α iff $Sz \leq \alpha$.

Two tests of the same significance level may be compared by their power. Fix a significance level $\alpha \in (0, 1)$. Consider two tests for the hypotheses in (1), each with power function $\gamma_1(\theta)$ and $\gamma_2(\theta)$. Both tests are of significance level α, i.e., $\max_{\theta \in \Theta_0} \gamma_1(\theta) \leq \alpha$ and

$\max_{\theta\in\Theta_0}\gamma_2(\theta)\leq\alpha$. The first test is **uniformly more powerful** than the second test iff

$$\begin{aligned}&\gamma_1(\theta)\geq\gamma_2(\theta)\text{ for all }\theta\in\Theta_1\text{ and}\\&\gamma_1(\theta)>\gamma_2(\theta)\text{ for some }\theta\in\Theta_1.\end{aligned}\tag{4}$$

Note that the "uniform" means uniform over $\theta\in\Theta_1$.

Claim 1. If a test of significance level α has size $Sz_1<\alpha$, there must be another test of significance level α that is uniformly more powerful.

Proof. We show this claim by construction. Suppose the first test rejects iff $T_1(\mathbb{X})>c$ for some test statistic $T_1(\mathbb{X})$ and critical value c. Introduce an auxiliary random variable ε. Conditional on $T_1(\mathbb{X})>c$, $\varepsilon=1$; and conditional on $T_1(\mathbb{X})\leq c$, $\varepsilon\in Bern(p)$ with $p=(1-\alpha)/(1-Sz_1)$. Let the test statistic of the new test be: $T_2(\mathbb{X},\varepsilon)=\varepsilon T_1(\mathbb{X})+(1-\varepsilon)(c+1)$. Let the new test's critical value be also c. Then the new test is of level α because:

$$\begin{aligned}Sz_2&=\max_{\theta\in\Theta_0}\Pr{}_\theta(T_2(\mathbb{X},\varepsilon)>c)\\&=\max_{\theta\in\Theta_0}[\Pr{}_\theta(T_2(\mathbb{X},\varepsilon)>c,\varepsilon=1)+\Pr{}_\theta(T_2(\mathbb{X},\varepsilon)>c,\varepsilon=0)]\\&=\max_{\theta\in\Theta_0}[\Pr{}_\theta(T_1(\mathbb{X})>c,\varepsilon=1)+\Pr{}_\theta(c+1>c,\varepsilon=0)]\\&=\max_{\theta\in\Theta_0}[\Pr{}_\theta(T_1(\mathbb{X})>c)+\alpha-Sz_1]\\&=\max_{\theta\in\Theta_0}\Pr{}_\theta(T_1(\mathbb{X})>c)+\alpha-Sz_1\\&=Sz_1+\alpha-Sz_1\\&=\alpha.\end{aligned}\tag{5}$$

The new test is uniformly more powerful than the first test because for any $\theta\in\Theta_1$ (in fact, in this example, for any $\theta\in\Theta$),

$$\begin{aligned}\gamma_2(\theta):&=\Pr{}_\theta(T_2(\mathbb{X},\varepsilon)>c)\\&=\Pr{}_\theta(T_1(\mathbb{X})>c)+\alpha-Sz_1\\&>\Pr{}_\theta(T_1(\mathbb{X})>c)\\&=:\gamma_1(\theta).\end{aligned}\tag{6}$$

Thus, the claim is proved. □

Remark. (1) If a test of significance level α has size strictly less than α, we say this test is conservative. A conservative test can be easily improved upon if we know how much smaller

Sz is than our desired significance level. The proof above gives us a way to do so, through a "randomized test". The new test in the proof above is called a "randomized test" because it uses some extra random variable that we generate (not from the data).

(2) The claim above does NOT imply that a conservative test is necessarily less powerful than a non-conservative test (i.e. one with $Sz = \alpha$). For example, the competely randomized test (that rejects H_0 iff a Bernoulli$(1-\alpha)$ independent of the data returns 1) is non-conservative, but its power is α, often much smaller than any test that uses the sample information in any reasonable way. In difficult testing situations, potentially conservative tests are used because it is difficult to figure out what Sz is exactly but it is easier to make (prove) $Sz \leq \alpha$.

A test of significance level α is said to be **uniformly most powerful (UMP)** if it is uniformly more powerful than any other test of significance level α.

A corollary of the claim above is that: a conservative test cannot be UMP.

UMP tests do not always exist. We will discuss several situations in which they do exist.

Case 1. Both H_0 and H_1 are simple hypotheses:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1\theta = \theta_1. \tag{7}$$

Lemma (Neyman-Pearson)**.** *Consider a random n-sample $\mathbb{X}$ from a distribution with density $f_X(x, \theta)$. For the H_0 vs. H_1 above, consider a test with rejection region C_R satisfy*

(1) any $\mathbb{X}$ such that $f_{\mathbb{X}}(\mathbb{X}, \theta_0)/f_{\mathbb{X}}(\mathbb{X}, \theta_1) < k$ belongs to C_R,

(2) any $\mathbb{X}$ such that $f_{\mathbb{X}}(\mathbb{X}, \theta_0)/f_{\mathbb{X}}(\mathbb{X}, \theta_1) > k$ does not belong to C_R, and

(3) $\Pr_{\theta_0}(\mathbb{X} \in C_R) = \alpha$.

Then the test is a UMP test of significance level α.

Proof. (in the book) □

Special case: If $\Pr_{\theta_0}(f_{\mathbb{X}}(\mathbb{X}, \theta_0)/f_{\mathbb{X}}(\mathbb{X}, \theta_1) = k) = 0$ at $k = k(\alpha)$ – the α quantile of $: f_{\mathbb{X}}(\mathbb{X}, \theta_0)/f_{\mathbb{X}}(\mathbb{X}, \theta_1)$, then the UMP rejection region is of the form: $C_R = \{\mathbb{X} : f_{\mathbb{X}}(\mathbb{X}, \theta_0)/f_{\mathbb{X}}(\mathbb{X}, \theta_1) < k(\alpha)\}$. Otherwise, one may need the UMP test to be a randomized test.

Example 1. $X \sim N(\theta, 1)$ and $\theta_1 > \theta_0$. Then

$$f_{\mathbb{X}}(x_1, ..., x_n, \theta) = \times_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2}.$$

The likelihood ratio is

$$\begin{aligned}
\frac{f_{\mathbb{X}}(\mathbb{X}, \theta_0)}{f_{\mathbb{X}}(\mathbb{X}, \theta_1)} &= \times_{i=1}^{n} e^{-(X_i-\theta_0)^2/2+(X_i-\theta_1)^2/2} \\
&= \times_{i=1}^{n} e^{(-2(\theta_1-\theta_0)X_i+\theta_1^2-\theta_0^2)/2} \\
&= e^{\sum_{i=1}^{n}(-(\theta_1-\theta_0)X_i+(\theta_1^2-\theta_0^2)/2)} \\
&= e^{(n/2)(\theta_1^2-\theta_0^2)-n(\theta_1-\theta_0)\bar{X}_n}. \qquad (8)
\end{aligned}$$

Under H_0, it is a continuous and strictly monotone function of a continuous random variable ($\bar{X}_n \sim N(\theta_0, 1/n)$). Thus, it satisfies the special case: for any $k > 0$,

$$\begin{aligned}
\Pr{}_{\theta_0}(f_{\mathbb{X}}(\mathbb{X}, \theta_0)/f_{\mathbb{X}}(\mathbb{X}, \theta_1) = k) &= \Pr{}_{\theta_0}((n/2)(\theta_1^2 - \theta_0^2) - n(\theta_1 - \theta_0)\bar{X}_n = \ln(k)) \\
&= \Pr{}_{\theta_0}(\bar{X}_n = 0.5(\theta_1 + \theta_0) - \ln(k)/(n(\theta_1 - \theta_0))) \\
&= 0. \qquad (9)
\end{aligned}$$

Thus, we just need to find $k(\alpha)$ in order to get the UMP test. $k(\alpha)$ should satisfy:

$$\Pr{}_{\theta_0}(f_{\mathbb{X}}(\mathbb{X}, \theta_0)/f_{\mathbb{X}}(\mathbb{X}, \theta_1) < k(\alpha)) = \alpha.$$

Thus, $k(\alpha)$ solves the following equation,

$$\begin{aligned}
\alpha &= \Pr{}_{\theta_0}((n/2)(\theta_1^2 - \theta_0^2) - n(\theta_1 - \theta_0)\bar{X}_n < \ln(k(\alpha))) \\
&= \Pr{}_{\theta_0}(\bar{X}_n > 0.5(\theta_1 + \theta_0) - \ln(k(\alpha))/(n(\theta_1 - \theta_0))) \\
&= \Pr{}_{\theta_0}(\sqrt{n}(\bar{X}_n - \theta_0) > 0.5(\theta_1 - \theta_0)\sqrt{n} - \ln(k(\alpha))/(\sqrt{n}(\theta_1 - \theta_0))) \\
&= 1 - \Phi(0.5(\theta_1 - \theta_0)\sqrt{n} - \ln(k(\alpha))/(\sqrt{n}(\theta_1 - \theta_0))),
\end{aligned}$$

We could solve the equaltion and get a closed form for $k(\alpha)$, but it is more useful to realize that, for $k(\alpha)$ that satisfy the above equation,

$$f_{\mathbb{X}}(\mathbb{X}, \theta_0)/f_{\mathbb{X}}(\mathbb{X}, \theta_1) < k(\alpha)$$

is equivalent to $\sqrt{n}(\bar{X}_n - \theta_0) > z_\alpha$, where z_α is the $1 - \alpha$ quantile of $N(0, 1)$. This tells

us that the UMP rejection region is of the form:

$$C_R = \{\mathbb{X} : t_n := \sqrt{n}(\bar{X}_n - \theta_0) > z_\alpha\}. \tag{10}$$

The t_n is called the t-statistic and the test is the t-test.

Notice that the UMP test in this example only depends on the null value θ_0, not on the alternative value θ_1. This suggests, C_R is also the UMP test for the following simple null against composite alternative hypotheses:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta > \theta_0. \tag{11}$$

Case 2. the null is simple and the alternative is composite.

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \in \Theta/\{\theta_0\}. \tag{12}$$

Lemma. *Consider a random n-sample $\mathbb{X}$ from a distribution with density $f_X(x, \theta)$. Assume that for any $\theta \in \Theta$, the likelihood ratio $f_{\mathbb{X}}(\mathbb{X}, \theta_0)/f_{\mathbb{X}}(\mathbb{X}, \theta_1)$ is an increasing function of a statistic $T(\mathbb{X})$. Then, for the H_0 vs. H_1 above, consider a test with rejection region C_R satisfy*

(1) any $\mathbb{X}$ such that $T(\mathbb{X}) < k$ belongs to C_R,

(2) any $\mathbb{X}$ such that $T(\mathbb{X}) > k$ does not belong to C_R, and

(3) $\Pr_{\theta_0}(\mathbb{X} \in C_R) = \alpha$.

Then the test is a UMP test of significance level α.

The additional assumption in the above Lemma relative to the Neyman-Pearson lemma is the "monotone likelihood ratio" assumption. The normal example above satisfy this assumption: the likelihood ratio is an increasing function of $-\sqrt{n}(\bar{X}_n - \theta_0)$.

(Think: what if we'd like to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$. in the normal example)?

Case 3. both the null and the alternative are composite. We only consider the one-dimensional example:

$$H_0 : \theta < \theta_0 \text{ vs. } H_1 : \theta \geq \theta_0. \tag{13}$$

Claim. If a test, C_R, of significance level α is UMP for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \geq \theta_0$ and $\inf_{\theta < \theta_0} \Pr_\theta(\mathbb{X} \in C_R) \leq \alpha$, then the test C_R is also a UMP test for $H_0 : \theta < \theta_0$ vs. $H_1 : \theta \geq \theta_0$ of significance level α.

Proof. (exercise) □

The Neyman-Pearson Lemma and its extensions motivates the use of the likelihood ratio test: to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, one can simply let the rejection region be the set of $\mathbb{X}$ that satisfies:

$$\frac{\max_{\theta\in\Theta_0} f_{\mathbb{X}}(\mathbb{X},\theta)}{\max_{\theta\in\Theta} f_{\mathbb{X}}(\mathbb{X},\theta)} < k(\alpha),$$

where $k(\alpha)$ is the $1-\alpha$ quantile of $\frac{\max_{\theta\in\Theta_0} f_{\mathbb{X}}(\mathbb{X},\theta)}{\max_{\theta\in\Theta} f_{\mathbb{X}}(\mathbb{X},\theta)}$. (Verify) In the three cases above, assuming $\frac{\max_{\theta\in\Theta_0} f_{\mathbb{X}}(\mathbb{X},\theta)}{\max_{\theta\in\Theta} f_{\mathbb{X}}(\mathbb{X},\theta)}$'s cdf is continuous at its $1-\alpha$ quantile, the likelihood ratio test is UMP.

The likelihood ratio test is often used in problems where the UMP test does not exist. It is relatively simple and has reasonably good power (though not necessarily UMP).

Typically, one uses the logarithm of the likelihood ratio (log-likelihood ratio) instead of the likelihood ratio itself to define the rejection region:

$$LR_n := \max_{\theta\in\Theta_0} \ln(f_{\mathbb{X}}(\mathbb{X},\theta)) - \max_{\theta\in\Theta} \ln(f_{\mathbb{X}}(\mathbb{X},\theta)) < c(\alpha),$$

where $c(\alpha)$ is the $1-\alpha$ quantile of LR_n.

In most non-normal settings, the exact distribution of $\max_{\theta\in\Theta_0} \ln(f_{\mathbb{X}}(\mathbb{X},\theta)) - \max_{\theta\in\Theta} \ln(f_{\mathbb{X}}(\mathbb{X},\theta))$ is too complicated and the quantile of it is too hard to find. But under fairly general conditions (for Θ_0 of certain shape), one can show $-2LR_n$ converges in distribution to a χ^2 random variable of a known degree of freedom under any $\theta \in \Theta_0$. This leads to the asymptotic version of the likelihood ratio test, where the χ^2 quantiles are used as critical values. The size of the asymptotic version of the test is only asymptotically controlled:

$$\lim_{n\to\infty} \max_{\theta\in\Theta_0} \Pr{}_\theta(-2LR_n > \chi^2_{1-\alpha,df}) = \alpha, \tag{14}$$

where *df* is the appropriate degree of freedom.

Some more words about the Neyman-Pearson framework for hypothesis testing.

The type-I error and the type-II error are treated asymmetrically. We control the type-I error to on or below the significance level α (which typically is a small number). And given that the type-I error is under control, we try to make the type-II bigger. But the type-II error can still be very big (close to $1-\alpha$) even for the UMP test. This is why a "rejection" is a stronger conclusion than an "acceptance" of H_0. If the test (of significance level α) rejects H_0, the probability that it is making an error is at most α. If the test accepts H_0, the maximum probability that it is making an error $\sup_{\theta\in\Theta_1}(1-\gamma(\theta))$, which could be big. For this reason, it is argued that one should use the hypothesis that one wants to reject as the null hypothesis and the opposite of it as the alternative hypothesis.

However because the type-I error and the type-II error are treated asymmetrically, H_0 and H_1 cannot be specified in any way we want. For example, Suppose one wants to argue that $\theta = 0$ (i.e. to reject the hypothesis that $\theta \neq 0$) and tries to specifies $H_0 : \theta \neq 0$ vs. $H_1 : \theta = 0$. If $f_X(x, \theta)$ is continuous in θ, then any test of significance level α cannot have power greater than α. This is because the power function $\gamma(\theta)$ is continuous in θ and $Sz := \sup_{\theta \neq 0} \gamma(\theta) \leq \alpha$ implies that $\gamma(0) \leq \alpha$.As a result, the test rejects at most $100\alpha\%$ of the time even when H_1 is true.

As you may noticed, the significance level α is important in the N-P framework. However, the choice of α is somewhat arbitrary. Conventionally, people use $\alpha = 0.01$, $\alpha = 0.05$ or $\alpha = 0.1$. There is not much reason why these three values should be used, except that they look small and they look simple. Because different α implies different choices of the critical value (different rejection regions), two researchers using the same test statistic for the same hypotheses may reach different conclusion depending on this α. This inconvenience prompts people to use the "empirical significance level", or the "p-value" (probability value). The p-value is the smallest significance level at which the test rejects H_0. A smaller p-value is stronger evidence against H_0. In practice, the p-value is computed by:

$$p - value = \max_{\theta \in \Theta_0} \Pr{}_\theta(T(\mathbb{X}) > T(x_1, ..., x_n)),$$

where $(x_1, ..., x_n)$ is the observed value of $\mathbb{X}$.