36-700: Probability and Mathematical Statistics **Spring 2019**

Lecture 12: Neyman-Pearson and Wald Tests

Lecturer: Jing Lei

12.1 Review and Outline

In the last class we discussed hypothesis testing:

1. Basic setup, the null and alternate hypothesis.
2. Construction of tests.
3. The power function and the Neyman-Pearson paradigm.

Today we will discuss a general method to construct the *optimal* test for particular simple hypothesis testing problems. This test is called the Neyman-Pearson test or the likelihood ratio test. After this we will discuss other general ways to construct tests. The second half will follow the Wasserman book.

12.2 The Neyman Pearson test

12.2.1 Simple versus simple hypothesis tests

The setting we will be interested in is where both the null and alternate hypotheses are simple hypotheses, i.e., there are two parameters θ_0 and θ_1 such that:

$$H_0 : \ \theta = \theta_0$$
$$H_1 : \ \theta = \theta_1.$$

We will denote the null density as f_0 and the alternate density as f_1.

12.2.2 The power of a statistical test

The power of a test is just the probability of correctly rejecting the null, i.e., the probability of rejecting the null when the alternate is true. To remind you the size of a test is the probability of incorrectly rejecting the null – this is the quantity we usually control at α.

We will also associate every test with a test function ϕ. We will keep things simple: for us tests are deterministic, i.e. once you see the data you either reject the null or accept it (you do not randomize this decision). Recall, that R is the rejection region for our test, i.e. if $x \in R$ we reject the null and otherwise we accept it. When our actions are deterministic the test function for a test T:

$$\phi(x) = \begin{cases} 1, & \text{if } x \in R \\ 0, & \text{if } x \notin R. \end{cases}$$

So we can write the power of a test as:

$$\int_x \phi(x) f_1(x) dx \quad \text{and the size as} \quad \int_x \phi(x) f_0(x) dx.$$

12.2.3 The Neyman Pearson test

The Neyman Pearson test statistic is to take the likelihood ratio:

$$\Lambda(x) = \frac{L(x;\theta_0)}{L(x;\theta_1)} = \frac{f_0(x)}{f_1(x)}$$

and to reject if this value is small. Again, we will compute the precise cut-off by controlling the probability of making a Type I error. That is we select a threshold t^* such that:

$$\mathbb{P}_0(\Lambda(X) \leq t^*) = \alpha.$$

One nice thing about this is that it is a "general recipe" for doing a hypothesis test. The drawback of course is that it only applies to the restricted class of simple versus simple tests.

The Neyman-Pearson test, despite its restricted applicability is a very important conceptual contribution. When it is applicable it is an optimal test. This is often called the Neyman-Pearson Lemma, and we provide a proof of it below.

12.2.4 The Neyman-Pearson Lemma

The Neyman-Pearson Lemma says that the NP test, is the most powerful test of size α. This means that if we have *any other test* that controls the Type I error rate at α, then its power is at most the power of the NP test.

Proof: Let us denote the test function of the NP test as ϕ_{NP} and the test function of any other test we want to compare against as ϕ_A.

To prove the NP Lemma, we will first argue that the following is true:

$$\int_x \underbrace{(\phi_{NP}(x) - \phi_A(x))}_{T_1} \underbrace{\left(f_1(x) - \frac{f_0(x)}{t^*}\right)}_{T_2} dx \geq 0.$$

To see this we can just consider some cases:

1. If both tests reject or if both tests accept then the inequality is clearly true since the LHS is 0.
2. If NP rejects, and the test A accepts then $\phi_{NP}(x) = 1$, and $\phi_A(x) = 0$, so $T_1 \geq 0$. Since the NP test rejected the null we know that:

   $$\frac{f_0(x)}{f_1(x)} \leq t^*,$$

   so that $T_2 \geq 0$. So the inequality is true in this case.
3. If NP accepts and the test A rejects then both T_1 and T_2 are negative so the inequality is also true in this case.

So we can see that for every x, $T_1 \times T_2 \geq 0$ so it is true when we integrate over x. Now, we can rearrange this inequality to see that:

$$\begin{aligned}
\int_x (\phi_{NP}(x) - \phi_A(x)) f_1(x) dx &\geq \frac{1}{t^*} \int_x (\phi_{NP}(x) - \phi_A(x)) f_0(x) dx \\
&= \frac{1}{t^*} \left(\underbrace{\int_x \phi_{NP}(x) f_0(x) dx}_{=\alpha} - \underbrace{\int_x \phi_A(x) f_0(x) dx}_{\leq \alpha} \right) \\
&\geq 0.
\end{aligned}$$

This proves the NP lemma, i.e. that the power of the NP test is larger than the power of any other test.

12.3 The Wald Test

The Wald test considers hypothesis testing problems of the form:

$$\begin{aligned}
H_0 &: \ \theta = \theta_0 \\
H_1 &: \ \theta \neq \theta_0,
\end{aligned}$$

although it is much more broadly applicable.

Wald basically suggested that we use an estimator that is asymptotically normal (under the null), i.e. for example we could use the MLE. We have seen that under the null hypothesis the MLE typically satisfies:

$$\sqrt{nI(\theta_0)}(\widehat{\theta} - \theta_0) \to N(0, 1),$$

in distribution as $n \to \infty$. So an idea would be to use this as our test statistic, i.e.:

$$T_n = \sqrt{nI(\theta_0)}(\widehat{\theta} - \theta_0),$$

and then construct the rejection threshold the same way we did in the last lecture. As in point estimation we could also estimate the Fisher information and use:

$$T_n = \sqrt{nI(\widehat{\theta})}(\widehat{\theta} - \theta_0).$$

This statistic will be more meaningful under the alternate since it will still have a standard normal distribution. We reject the null hypothesis if:

$$|T_n| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

It is easy to see that asymptotically this is a correct size-α test.

Example: Suppose that we were testing the null $H_0 : p = p_0$ in a Bernoulli problem. We would use:

$$T_n = \frac{\widehat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}},$$

and reject the null if $|T_n| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

12.3.1 Wald test for two sample mean comparison

Now lets consider the Wald test for comparing two Bernoullis. Suppose that we have two prediction algorithms that we would like to compare. There two typical scenarios.

Scenario 1: Suppose first, that we test the two algorithms by evaluating them on two different test sets: Algorithm 1 on a test set of size m and Algorithm 2 on a test set of size n. Let X be the number of correct predictions made by Algorithm 1 and Y be the number of correct predictions made by Algorithm 2. Then $X \sim \text{Binomial}(m, p_1)$ and $Y \sim \text{Binomial}(n, p_2)$ where p_1 and p_2 are the error rates of the first and second Algorithm respectively. Defining $\delta = p_1 - p_2$, our hypotheses are:

$$\begin{aligned} H_0 : &\quad \delta = 0 \\ H_1 : &\quad \delta \neq 0. \end{aligned}$$

To do a Wald test, we compute the MLEs $\widehat{p}_1$ and $\widehat{p}_2$, and compute $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$. We can also estimate the variance of our estimator in the usual way:

$$\mathrm{Var}(\widehat{\delta}) = \frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}.$$

So our Wald test would reject the null hypothesis if

$$\frac{|\widehat{\delta}|}{\sqrt{\mathrm{Var}(\widehat{\delta})}} \geq \Phi^{-1}(1-\alpha/2).$$

Scenario 2. Another typical setting is when we have a single data set on which we evaluate both algorithms. In this case the two predictions are not independent so we cannot directly compute the variance of $\widehat{\delta}$. You can try to do this and observe that it will depend on the covariance between the two algorithms which you will have to estimate.

A convenient way to test this hypothesis is via a paired comparison test. In this case, one can imagine testing both Algorithms on the data but recording, the difference $D_i = X_i - Y_i$, i.e.

1. $D_i = 0$ if both algorithms have the same prediction.
2. $D_i = 1$ if Algorithm 1 is correct while 2 is wrong.
3. $D_i = -1$ if Algorithm 2 is correct while 1 is wrong.

In this case you can observe that:

$$\delta = \mathbb{E}(X) - \mathbb{E}(Y) = \mathbb{E}(X - Y) = \mathbb{E}(D).$$

Our test statistic is simply:

$$T_n = \frac{\frac{1}{n}\sum_{i=1}^n D_i}{\sqrt{\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n D_i)}},$$

and the variance is:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n D_i\right) = \frac{1}{n^2}\sum_{i=1}^n \left(D_i - \frac{1}{n}\sum_{i=1}^n D_i\right)^2.$$

As usual the Wald test would reject the null hypothesis if:

$$|T_n| \geq \Phi^{-1}(1-\alpha/2).$$

12.3.2 The power of the Wald test

Suppose that we use the statistic:

$$T_n = \sqrt{nI(\widehat{\theta})}(\widehat{\theta} - \theta_0),$$

and that the true value of the parameter is $\theta_1 \neq \theta_0$ then the probability that the Wald test rejects the null hypothesis is roughly:

$$1 - \Phi\left(\sqrt{nI(\theta_1)}(\theta_0 - \theta_1) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) + \Phi\left(\sqrt{nI(\theta_1)}(\theta_0 - \theta_1) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right).$$

We will not prove this but it follows just from some simple algebra on the power function evaluated at θ_1. There are some aspects to notice:

1. If the difference between θ_0 and θ_1 is very small the power will tend to 0.
2. As $n \to \infty$ the two Φ terms will approach either 0 or 1, and so the power will approach 1.
3. As a rule of thumb the Wald test will have non-trivial power if $|\theta_0 - \theta_1| \gg \frac{1}{\sqrt{nI(\theta_1)}}$.