36-705: Intermediate Statistics **Fall 2017**

Lecture 20: October 17

Lecturer: Siva Balakrishnan

In the last class we showed that the Neyman-Pearson test is optimal for testing simple versus simple hypothesis tests. Today we will develop some generalizations and tests that are useful in other more complex settings.

20.1 The Wald Test

When we are testing a simple null hypothesis against a possibly composite alternative, the NP test is no longer applicable and a general alternative is to use the Wald test.

We are interested in testing the hypotheses in a parametric model:

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

The Wald test most generally is based on an asymptotically normal estimator, i.e. we suppose that we have access to an estimator $\widehat{\theta}$ which under the null satisfies the property that:

$$\widehat{\theta} \xrightarrow{d} N(\theta_0, \sigma_0^2),$$

where σ_0^2 is the variance of the estimator under the null. The canonical example is when $\widehat{\theta}$ is taken to be the MLE.

In this case, we could consider the statistic:

$$T_n = \frac{\widehat{\theta} - \theta_0}{\sigma_0},$$

or if σ_0 is not known we can plug-in an estimate to obtain the statistic,

$$T_n = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma_0}}.$$

Under the null $T_n \xrightarrow{d} N(0,1)$, so we simply reject the null if: $T_n \geq z_{\alpha/2}$. This controls the Type-I error only asymptotically (i.e. only if $n \to \infty$) but this is relatively standard in applications.

Example: Suppose we considered the problem of testing the parameter of a Bernoulli, i.e. we observe $X_1, \ldots, X_n \sim \text{Ber}(p)$, and the null is that $p = p_0$. Defining $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. A Wald test could be constructed based on the statistic:

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

which has an asymptotic $N(0,1)$ distribution. An alternative would be to use a slightly different estimated standard deviation, i.e. to define,

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

Observe that this alternative test statistic also has an asymptotically standard normal distribution under the null. Its behaviour under the alternate is a bit more pleasant as we will see.

20.1.1 Power of the Wald Test

To get some idea of what happens under the alternate, suppose we are in some situation where the MLE has "standard asymptotics", i.e. $\widehat{\theta} - \theta \xrightarrow{d} N(0, 1/(nI_1(\theta)))$. Suppose that we use the statistic:

$$T_n = \sqrt{nI_1(\widehat{\theta})}(\widehat{\theta} - \theta_0),$$

and that the true value of the parameter is $\theta_1 \neq \theta_0$. Let us define:

$$\Delta = \sqrt{nI_1(\theta_1)}(\theta_0 - \theta_1),$$

then the probability that the Wald test rejects the null hypothesis is asymptotically:

$$1 - \Phi\left(\Delta + z_{\alpha/2}\right) + \Phi\left(\Delta - z_{\alpha/2}\right).$$

You will prove this on your HW (it is some simple re-arrangement, similar to what we have done previously when computing the power function in a Gaussian model). There are some aspects to notice:

1. If the difference between θ_0 and θ_1 is very small the power will tend to α, i.e. if $\Delta \approx 0$ then the test will have trivial power.
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 $|\theta_0 - \theta_1| \gg \frac{1}{\sqrt{nI_1(\theta_1)}}$.

20.2 Likelihood Ratio Test (LRT)

To test composite versus composite hypotheses the general method is to use something called the (generalized) likelihood ratio test.

We want to test:

$$\begin{aligned} H_0 : &\quad \theta \in \Theta_0 \\ H_1 : &\quad \theta \notin \Theta_0. \end{aligned}$$

This test is simple: reject H_0 if $\lambda(X_1, \ldots, X_n) \leq c$ where

$$\lambda(X_1, \ldots, X_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})}$$

where $\widehat{\theta}_0$ maximizes $L(\theta)$ subject to $\theta \in \Theta_0$.

Example 20.1 *$X_1, \ldots, X_n \sim N(\theta, 1)$. Suppose*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0.$$

After some algebra,

$$\lambda = \exp\left\{-\frac{n}{2}(\overline{X}_n - \theta_0)^2\right\}.$$

So

$$R = \{x : \ \lambda \leq c\} = \{x : \ |\overline{X} - \theta_0| \geq c'\}$$

where $c' = \sqrt{-2 \log c/n}$. Choosing c' to make this level α gives: reject if $|T_n| > z_{\alpha/2}$ where $T_n = \sqrt{n}(\overline{X} - \theta_0)$ which is the test we constructed before.

Example 20.2 *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. Suppose*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0.$$

Then

$$\lambda(x_1, \ldots, x_n) = \frac{L(\theta_0, \widehat{\sigma}_0)}{L(\widehat{\theta}, \widehat{\sigma})}$$

where $\widehat{\sigma}_0$ maximizes the likelihood subject to $\theta = \theta_0$.

Exercise: *Show that $\lambda(x_1, \ldots, x_n) < c$ corresponds to rejecting when $|T_n| > k$ for some constant k where*

$$T_n = \frac{\overline{X}_n - \theta_0}{S/\sqrt{n}}.$$

Under H_0, T_n has a t-distribution with $n-1$ degrees of freedom. So the final test is: reject H_0 if

$$|T_n| > t_{n-1,\alpha/2}.$$

This is called Student's t-test. It was invented by William Gosset working at Guiness Breweries and writing under the pseudonym Student.

We can simplify the LRT by using an asymptotic approximation. This fact that the LRT generally has a simple asymptotic approximation is known as *Wilks' phenomenon.* First, some notation:

Notation: Let $W \sim \chi^2_p$. Define $\chi^2_{p,\alpha}$ by

$$P(W > \chi^2_{p,\alpha}) = \alpha.$$

We let $\ell(\theta)$ denote the log-likelihood in what follows.

Theorem 20.3 *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ where $\theta \in \mathbb{R}$. Under H_0,*

$$-2\log\lambda(X_1,\ldots,X_n) \rightsquigarrow \chi^2_1.$$

Hence, if we let $T_n = -2\log\lambda(X^n)$ then

$$P_{\theta_0}(T_n > \chi^2_{1,\alpha}) \to \alpha$$

as $n \to \infty$.

Proof: Using a Taylor expansion:

$$\ell(\theta) \approx \ell(\widehat{\theta}) + \ell'(\widehat{\theta})(\theta - \widehat{\theta}) + \ell''(\widehat{\theta})\frac{(\theta-\widehat{\theta})^2}{2} = \ell(\widehat{\theta}) + \ell''(\widehat{\theta})\frac{(\theta-\widehat{\theta})^2}{2}$$

and so

$$\begin{aligned} -2\log\lambda(x_1,\ldots,x_n) &= 2\ell(\widehat{\theta}) - 2\ell(\theta_0) \\ &\approx 2\ell(\widehat{\theta}) - 2\ell(\widehat{\theta}) - \ell''(\widehat{\theta})(\theta_0-\widehat{\theta})^2 = -\ell''(\widehat{\theta})(\theta_0-\widehat{\theta})^2 \\ &= \frac{-\frac{1}{n}\ell''(\widehat{\theta})}{I_1(\theta_0)}(\sqrt{nI_1(\theta_0)}(\widehat{\theta}-\theta_0))^2 = A_n \times B_n. \end{aligned}$$

Now $A_n \xrightarrow{p} 1$ by the WLLN and $\sqrt{B_n} \rightsquigarrow N(0,1)$. The result follows by Slutsky's theorem. ■

Example 20.4 *$X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. We want to test $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$. Then*

$$-2 \log \lambda(x^n) = 2n[(\lambda_0 - \widehat{\lambda}) - \widehat{\lambda} \log(\lambda_0/\widehat{\lambda})].$$

We reject H_0 when $-2 \log \lambda(x^n) > \chi^2_{1,\alpha}$.

Now suppose that $\theta = (\theta_1, \ldots, \theta_k)$. Suppose that $H_0 : \theta \in \Theta_0$ fixes some of the parameters. Then, under conditions,

$$T_n = -2 \log \lambda(X_1, \ldots, X_n) \rightsquigarrow \chi^2_\nu$$

where

$$\nu = \dim(\Theta) - \dim(\Theta_0).$$

Therefore, an asymptotic level α test is: reject H_0 when $T_n > \chi^2_{\nu,\alpha}$.

Example 20.5 *Consider a multinomial with $\theta = (p_1, \ldots, p_5)$. So*

$$L(\theta) = p_1^{y_1} \cdots p_5^{y_5}.$$

Suppose we want to test

$$H_0 : p_1 = p_2 = p_3 \quad \text{and} \quad p_4 = p_5$$

versus the alternative that H_0 is false. In this case

$$\nu = 4 - 1 = 3.$$

The LRT test statistic is

$$\lambda(x_1, \ldots, x_n) = \frac{\prod_{j=1}^5 \widehat{p}_{0j}^{Y_j}}{\prod_{j=1}^5 \widehat{p}_j^{Y_j}}$$

where $\widehat{p}_j = Y_j/n$, $\widehat{p}_{01} = \widehat{p}_{02} = \widehat{p}_{03} = (Y_1 + Y_2 + Y_3)/n$, $\widehat{p}_{04} = \widehat{p}_{05} = (1 - 3\widehat{p}_{01})/2$. Now we reject H_0 if $-2\lambda(X_1, \ldots, X_n) > \chi^2_{3,\alpha}$. □

20.3 p-values

When we test at a given level α we will reject or not reject. It is useful to summarize what levels we would reject at and what levels we would not reject at.

The p-value is the smallest α at which we would reject H_0.

In other words, we reject at all $\alpha \geq p$. So, if the pvalue is 0.03, then we would reject at $\alpha = 0.05$ but not at $\alpha = 0.01$.

Hence, to test at level α, we reject when $p < \alpha$.

Theorem 20.6 *Suppose we have a test of the form: reject when $T(X_1, \ldots, X_n) > c$. Then the p-value is*

$$p = \sup_{\theta \in \Theta_0} P_\theta(T_n(X_1, \ldots, X_n) \geq T_n(x_1, \ldots, x_n))$$

where $x_1, \ldots, x_n$ are the observed data and $X_1, \ldots, X_n \sim p_{\theta_0}$.

Example 20.7 *$X_1, \ldots, X_n \sim N(\theta, 1)$. Test that $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We reject when $|T_n|$ is large, where $T_n = \sqrt{n}(\overline{X}_n - \theta_0)$. Let t_n be the obsrved value of T_n. Let $Z \sim N(0, 1)$. Then,*

$$p = P_{\theta_0}\left(|\sqrt{n}(\overline{X}_n - \theta_0)| > t_n\right) = P(|Z| > t_n) = 2\Phi(-|t_n|).$$

Theorem 20.8 *Under H_0, $p \sim \text{Unif}(0, 1)$.*

Important. Note that p is NOT equal to $P(H_0|X_1, \ldots, X_n)$. The latter is a Bayesian quantity which we will discuss later.

20.4 Goodness-of-fit testing

There are many testing problems that go beyond the usual parameteric testing problems we have formulated so far. Since we may not get a chance to cover these in detail later on, I thought it might be useful to discuss them briefly here.

The most canonical non-parametric testing problem is called goodness-of-fit testing. Here given samples $X_1, \ldots, X_n \sim P$, we want to test:

$$\begin{aligned} H_0 : &\quad P = P_0 \\ H_1 : &\quad P \neq P_0, \end{aligned}$$

for some fixed, known distribution P_0.

As a hypothetical example, you collect some measurements from a light source, you believe that the number of particles per unit time should have a Poisson distribution with a certain rate parameter (the intensity), and want to test this hypothesis.

20.4.1 The χ^2 test

In the simplest setting, P_0 and P are multinomials on k categories, i.e. the null distribution just a vector of probabilities $(p_{01}, \ldots, p_{0k})$, with $p_{0i} \geq 0$, $\sum_i p_{0i} = 1$.

Given a sample $X_1, \ldots, X_n$ you can reduce it to a vector of counts $(Z_1, \ldots, Z_k)$ where Z_i is the number of times you observed the i-th category.

A natural test statistic in this case (you could also do the likelihood ratio test) is to consider:

$$T(X_1, \ldots, X_n) = \sum_{i=1}^{k} \frac{(Z_i - np_{0i})^2 - np_{0i}}{np_{0i}}.$$

On your HW you will show that asymptotically this test statistic, under the null, has a χ^2_{k-1} distribution. This is called Pearson's χ^2 test.

More generally, you could do perform any goodness-of-fit test by reducing to a multinomial test by binning, i.e. you define a sufficiently find partition of the domain, this induces a multinomial p_0 under the null which you then test using Pearson's test.

20.5 Two-sample Testing

Another popular hypothesis testing problem is the following: you observe $X_1, \ldots, X_{n_1} \sim P$ and $Y_1, \ldots, Y_{n_2} \sim Q$, and want to test if:

$$\begin{aligned} H_0 : &\quad P = Q \\ H_1 : &\quad P \neq Q. \end{aligned}$$

There are many popular ways of testing this (for instance, in the ML literature kernel-based tests are quite popular – search for "Maximum Mean Discrepancy" if you are curious).

Suppose again we considered the multinomial setting where P and Q are multinomials on k categories. Then there is a version of the χ^2 test that is commonly used. Let us define $(Z_1, \ldots, Z_k)$ and $(Z'_1, \ldots, Z'_k)$ to be the counts in the X and Y sample respectively. We can define for $i \in \{1, \ldots, k\}$,

$$\widehat{c}_i = \frac{Z_i + Z'_i}{n_1 + n_2}.$$

The two-sample χ^2 test is then:

$$T_n = \sum_{i=1}^{k} \left[\frac{(Z_i - n_1\widehat{c}_i)^2}{n_1\widehat{c}_i} + \frac{(Z'_i - n_2\widehat{c}_i)^2}{n_2\widehat{c}_i} \right].$$

This is a bit harder to see but under the null this statistic also has a χ^2_{k-1} distribution. For two-sample testing we can determine the cutoff in a different way without resorting to asymptotics. This is called a permutation test. We will explore this in the general (not multinomial) setting.

20.5.1 The Permutation Test

Suppose we have data

$$X_1, \ldots, X_n \sim F$$

and

$$Y_1, \ldots, Y_m \sim G.$$

We want to test:

$$H_0 : F = G \quad \text{versus} \quad H_1 : F \neq G.$$

Let

$$Z = (X_1, \ldots, X_n, Y_1, \ldots, Y_m).$$

Create labels

$$L = (\underbrace{1, \ldots, 1}_{n \text{ values}}, \underbrace{2, \ldots, 2}_{m \text{ values}}).$$

A test statistic can be written as a function of Z and L. For example, if

$$T = |\overline{X}_n - \overline{Y}_m|$$

then we can write

$$T = \left| \frac{\sum_{i=1}^N Z_i I(L_i = 1)}{\sum_{i=1}^N I(L_i = 1)} - \frac{\sum_{i=1}^N Z_i I(L_i = 2)}{\sum_{i=1}^N I(L_i = 2)} \right|$$

where $N = n + m$. So we write $T = g(L, Z)$.

Define

$$p = \frac{1}{N!} \sum_{\pi} I(g(L_\pi, Z) > g(L, Z))$$

where L_π is a permutation of the labels and the sum is over all permutations. Under H_0, permuting the labels does not change the distribution. In other words, $g(L, Z)$ has an equal chance of having any rank among all the permuted values. That is, under H_0, $\approx \text{Unif}(0, 1)$ and if we reject when $p < \alpha$, then we have a level α test.

Summing over all permutations is infeasible. But it suffices to use a random sample of permutations. So we do this:

1. Compute a random permutation of the labels and compute W. Do this K times giving values $T^{(1)}, \ldots, T^{(K)}$.

2. Compute the p-value

$$\frac{1}{K} \sum_{j=1}^K I(T^{(j)} > T).$$