Lecture 21: October 19

Lecturer: Siva Balakrishnan

21.1 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 $\theta \in \Theta_0$.

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 21.1 *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 21.2 *$X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. 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 21.3 *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 $-2\lambda(X_1, \ldots, X_n) > \chi^2_{3,\alpha}$. □

21.2 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 21.4 *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 21.5 *$X_1, \ldots, X_n \sim N(\theta, 1)$. 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,*

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

The p-value is a random variable. Under some assumptions that you will see in your HW the p-value will be uniformly distributed on $[0, 1]$ under the null.

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.

21.3 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.

21.3.1 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.

21.4 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.

A typical example is in a drug trial where one set of people are given a drug and the other set are given a placebo. We then would like to know if there is some difference in the outcomes of the two populations or if they are identically distributed.

There are various possible test statistics, but a common one is to use a difference in means:

$$T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = \left| \frac{1}{m} \sum_{i=1}^{m} X_i - \frac{1}{n} \sum_{i=1}^{n} Y_i \right|,$$

one could also standardize this statistic by its variance, or consider more complex test statistics based on signs and ranks. Let us denote the test statistic computed on the data we observed as T_{obs}.

In general, since we have not assumed anything about F_X and F_Y it is not easy to compute the distribution of our test statistic, and approximations (based on a CLT for instance) might be quite bad. The permutation test, gives a way to design an *exact* α level test without making any approximations.

The idea of the permutation test is simple. Define $N = m + n$ and consider all $N!$ permutations of the data $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$. For each permutation we could compute our test statistic T. Denote these as $T_1, \ldots, T_{N!}$.

The key observation is: **under the null hypothesis each value $T_1, \ldots, T_{N!}$ has the *same* distribution (even if we do not know what it is).**

Suppose we reject for large values of T. Then we could simply define the p-value as:

$$\text{p-value} = \frac{1}{N!} \sum_{i=1}^{N!} \mathbb{I}(T_i > T_{\text{obs}}).$$

It is important to note that this is an exact p-value, i.e. no asymptotic approximations are needed to show that rejecting the null when this p-value is less than α controls the Type I error at α. Here is a toy-example from the Wasserman book:

Example 2: Suppose we observe $(X_1, X_2, Y_1) = (1, 9, 3)$. Let $T(X_1, X_2, Y_1)$ be the difference in means, i.e. $T(X_1, X_2, Y_1) = 2$. The permutations are:

permutation	value of T
(1,9,3)	2
(9,1,3)	2
(1,3,9)	7
(3,1,9)	7
(3,9,1)	5
(9,3,1)	5

We could use this to calculate the p-value by counting how often we got a larger value than 2:

$$\text{p-value} = \frac{4}{6} = 0.66,$$

so most likely we would not reject the null hypothesis in this case. Typically, we do not calculate the exact p-value (although in principle we could) since evaluating $N!$ test statistics would take too long for large N. Instead we approximate the p-value by drawing a few random permutations and using them. This leads to the following algorithm for computing the p-value using a permutation test:

Algorithm for Permutation Test

1. Compute the observed value of the test statistic $t_{\text{obs}} = T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$.
2. Randomly permute the data. Compute the statistic again using the permuted data.
3. Repeat the previous step B times and let $T_1, \ldots, T_B$ denote the resulting values.
4. The approximate p-value is

$$\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

21.5 Multiple Testing

The problem of multiple testing is one that is fundamental to a lot of science. Typical modern scientific discovery does not proceed in a simple fashion where we have a single hypothesis that we would like to test.

A classical example is in the analysis of gene expression data. We measure the expression of tens of thousands of genes and we would like to know if any of them are associated with some phenotype (for example whether a person has a disease or not). Typically, the way this is done is that the scientist does tens of thousands of hypothesis tests, and then reports the associations that are significant, i.e. reports the tests where the null hypothesis was rejected.

This is very problematic:

Suppose we did 1000 hypothesis tests, and for each of them rejected the null when the p-value was less than $\alpha = 0.05$. How many times would you expect to falsely reject the null hypothesis?

The answer is we would expect to reject the null hypothesis 50 times. So we really cannot report all the discovered associations (rejections) as significant because we expect many false rejections.

The multiple testing problem is behind a lot of the "reproducibility crisis" of modern science. Many results that have been reported significant cannot be reproduced simply because they are false rejections. Too many false rejections come from doing multiple testing but not properly adjusting your tests to reflect the fact that many hypothesis tests are being done[1].

The basic question is how to we adjust our p-value cutoffs to account for the fact that multiple tests are being done.

21.5.1 The Family-Wise Error Rate

We first need to define what the error control we desire is. Recall, the Type I error controls the probability of falsely rejecting the null hypothesis. We have seen that in order to control the Type I error we can simply threshold the p-value, i.e rejecting the null if the p-value $\leq \alpha$ controls the Type I error at α.

One possibility (and we will discuss a different one in the next lecture) is that when a scientist does multiple tests we care about controlling the probability that we falsely reject *any* null hypothesis. This is called the Family-Wise Error Rate (FWER).

The FWER is the probability of falsely rejecting the null hypothesis even once amongst the multiple tests.

[1]Too many false rejections can also arise from tests that do not properly control the α level but this is usually easier to detect/fix.

The basic question is then: how do we control the FWER?

21.5.2 Sidak correction

Suppose we do d hypothesis tests, and want to control the FWER at α.

The Sidak correction says we reject any test if the p-value is smaller than:

$$\text{p-value} \leq 1 - (1-\alpha)^{1/d} = \alpha_t,$$

so we reject any test if its p-value is less than α_t.

The main result is that: if the p-values are all *independent* then the FWER $\leq \alpha$.

Proof: Suppose that all the null hypotheses are true (this is called the *global null*). You can easily see that if this is not the case you can simply ignore all the tests for which the null is false. The probability of falsely rejecting a fixed test is α_t, so we correctly fail to reject it with probability $1 - \alpha_t$.

Since the p-values are all independent the probability of falsely rejecting any null hypothesis is:

$$\text{FWER} = 1 - (1-\alpha_t)^d = \alpha.$$

21.5.3 Bonferroni correction

The main problem with the Sidak correction is that it requires the independence of p-values. This is unrealistic especially if you compute the test statistics for the different tests on the same set of data. The Bonferroni correction instead uses the union bound to avoid this assumption.

The Bonferroni correction says we reject any test if the p-value is smaller than:

$$\text{p-value} \leq \frac{\alpha}{d}.$$

The main result is that: The FWER $\leq \alpha$.

Proof: Suppose again that the global null is true. In this case,

$$\text{FWER} = \mathbb{P}\left(\bigcup_{i=1}^{d} \text{reject } H_{0i}\right) \leq \sum_{i=1}^{d} \mathbb{P}\left(\text{reject } H_{0i}\right) \leq \sum_{i=1}^{d} \frac{\alpha}{d} = \alpha,$$

where the first inequality follows from the union bound.

21.5.4 Something to think about

In the above discussion we assumed that there was a single scientist doing a bunch of tests so he could appropriately correct his procedure for the multiple testing problem.

One thing to ponder is really what error rate should we be controlling, i.e. maybe I am the editor of a journal, and I want to ensure that across all articles in my journal the FWER is $\leq \alpha$. Maybe I want this to be true across the entire field? Should I be adjusting my p-values for people in other disciplines? Sounds absurd but it actually makes sense if you think about each of these procedures and their implications for reproducibility.