Lecture Notes 15
Hypothesis Testing (Chapter 10)

1 Introduction

Let $X_1, \ldots, X_n \sim p_\theta(x)$. Suppose we we want to know if $\theta = \theta_0$ or not, where θ_0 is a specific value of θ. For example, if we are flipping a coin, we may want to know if the coin is fair; this corresponds to $p = 1/2$. If we are testing the effect of two drugs — whose means effects are θ_1 and θ_2 — we may be interested to know if there is no difference, which corresponds to $\theta_1 - \theta_2 = 0$.

We formalize this by stating a *null hypothesis* H_0 and an alternative hypothesis H_1. For example:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad \theta \neq \theta_0.$$

More generally, consider a parameter space Θ. We consider

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cap \Theta_1 = \emptyset$. If Θ_0 consists of a single point, we call this a *simple null hypothesis.* If Θ_0 consists of more than one point, we call this a *composite null hypothesis.*

Example 1 $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}. \quad \square$$

The question is not whether H_0 is true or false. The question is whether there is sufficient evidence to reject H_0, much like a court case. Our possible actions are: reject H_0 or retain (don't reject) H_0.

	Decision	
	Retain H_0	Reject H_0
H_0 true	$\surd$	Type I error (false positive)
H_1 true	Type II error (false negative)	$\surd$

Warning: Hypothesis testing should only be used when it is appropriate. Often times, people use hypothesis testing when it would be much more appropriate to use confidence intervals.

Notation: Let Φ be the cdf of a standard Normal random variable Z. For $0 < \alpha < 1$, let

$$z_\alpha = \Phi^{-1}(1-\alpha).$$

Hence,

$$P(Z > z_\alpha) = \alpha.$$

Also, $P(Z < -z_\alpha) = \alpha$. In these notes we sometimes write $p(x;\theta)$ instead of $p_\theta(x)$.

2 Constructing Tests

Hypothesis testing involves the following steps:

1. Choose a *test statistic* $T_n = T_n(X_1, \ldots, X_n)$.
2. Choose a rejection region R.
3. If $T_n \in R$ we reject H_0 otherwise we retain H_0.

Example 2 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *Suppose we test*

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}.$$

Let $T_n = n^{-1}\sum_{i=1}^n X_i$ *and* $R = \{x_1, \ldots, x_n : \ |T_n(x_1, \ldots, x_n) - 1/2| > \delta\}$. *So we reject* H_0 *if* $|T_n - 1/2| > \delta$.

We need to choose T and R so that the test has good statistical properties. We will consider the following tests:

1. The Neyman-Pearson Test
2. The Wald test
3. The Likelihood Ratio Test (LRT)
4. The permutation test.

Before we discuss these methods, we first need to talk about how we evaluate tests.

3 Error Rates and Power

Suppose we reject H_0 when $(X_1, \ldots, X_n) \in R$. Define the *power function* by

$$\beta(\theta) = P_\theta(X_1, \ldots, X_n \in R).$$

We want $\beta(\theta)$ to be small when $\theta \in \Theta_0$ and we want $\beta(\theta)$ to be large when $\theta \in \Theta_1$. The general strategy is:

1. Fix $\alpha \in [0, 1]$.

2. Now try to maximize $\beta(\theta)$ for $\theta \in \Theta_1$ subject to $\beta(\theta) \le \alpha$ for $\theta \in \Theta_0$.

We need the following definitions. A test is *size* α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \le \alpha.$$

Example 3 *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Suppose we test*

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

This is called a **one-sided alternative**. *Suppose we reject H_0 if $T_n > c$ where*

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

Then

$$\begin{aligned} \beta(\theta) &= P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > c\right) = P_\theta\left(\frac{\overline{X}_n - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \end{aligned}$$

where Φ is the cdf of a standard Normal and $Z \sim \Phi$. Now

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = 1 - \Phi(c).$$

To get a size α test, set $1 - \Phi(c) = \alpha$ so that

$$c = z_\alpha$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$. Our test is: reject H_0 when

$$T_n = \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > z_\alpha.$$

Example 4 *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Suppose*

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

This is called a **two-sided** *alternative. We will reject H_0 if $|T_n| > c$ where T_n is defined as before. Now*

$$\begin{aligned} \beta(\theta) &= P_\theta(T_n < -c) + P_\theta(T_n > c) \\ &= P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} < -c\right) + P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > c\right) \\ &= P\left(Z < -c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + \Phi\left(-c - \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \end{aligned}$$

since $\Phi(-x) = 1 - \Phi(x)$. *The size is*

$$\beta(\theta_0) = 2\Phi(-c).$$

To get a size α *test we set* $2\Phi(-c) = \alpha$ *so that* $c = -\Phi^{-1}(\alpha/2) = \Phi^{-1}(1-\alpha/2) = z_{\alpha/2}$. *The test is: reject* H_0 *when*

$$|T| = \left| \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}.$$

4 The Neyman-Pearson Test

(Not in the book.) Let $\mathcal{C}_\alpha$ denote all level α tests. A test in $\mathcal{C}_\alpha$ with power function β is **uniformly most powerful (UMP)** if the following holds: if β' is the power function of any other test in $\mathcal{C}_\alpha$ then $\beta(\theta) \leq \beta'(\theta)$ for all $\theta \in \Theta_1$.

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. (Simple null and simple alternative.)

Theorem 5 *Let* $L(\theta) = p(X_1, \ldots, X_n; \theta)$ *and*

$$T_n = \frac{L(\theta_1)}{L(\theta_0)}.$$

Suppose we reject H_0 *if* $T_n > k$ *where* k *is chosen so that*

$$P_{\theta_0}(X^n \in R) = \alpha.$$

This test is a UMP level α *test.*

The Neyman-Pearson test is quite limited because it can be used only for testing a simple null versus a simple alternative. So it does not get used in practice very often. But it is important from a conceptual point of view.

5 The Wald Test

Let

$$T_n = \frac{\widehat{\theta}_n - \theta_0}{\mathrm{se}}$$

where $\widehat{\theta}$ is an asymptotically Normal estimator and se is the estimated standard error of $\widehat{\theta}$ (or the standard error under H_0). Under H_0, $T_n \rightsquigarrow N(0,1)$. Hence, an asymptotic level α test is to reject when $|T_n| > z_{\alpha/2}$. That is

$$P_{\theta_0}(|T_n| > z_\alpha) \to \alpha.$$

For example, with Bernoulli data, to test $H_0 : p = p_0$,

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

You can also use

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

In other words, to compute the standard error, you can replace θ with an estimate $\widehat{\theta}$ or by the null value θ_0.

6 The Likelihood Ratio Test (LRT)

This test is simple: reject H_0 if $\lambda(x_1, \ldots, x_n) \le c$ 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 6 $X_1, \ldots, X_n \sim N(\theta, 1)$. *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

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

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 7 $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. *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. First, some notation:

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

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

Theorem 8 *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

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

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

Example 9 $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. *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 10 *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

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

where $\widehat{p}_j = Y_j/n$, $\widehat{p}_{10} = \widehat{p}_{20} = \widehat{p}_{30} = (Y_1 + Y_2 + Y_3)/n$, $\widehat{p}_{40} = \widehat{p}_{50} = (1 - 3\widehat{p}_{10})/2$. *These calculations are on p 491. Make sure you understand them. Now we reject* H_0 *if* $-2\lambda(X_1, \ldots, X_n) > \chi^2_{3,\alpha}$. $\square$

7 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 woud 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 when $p < \alpha$.

Theorem 11 *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 12 $X_1, \ldots, X_n \sim N(\theta, 1)$*. 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 13 *Under* H_0, $p \sim \mathrm{Unif}(0, 1)$.

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.

8 The Permutation Test

This is a very cool test. It is distribution free and it does not involve any asymptotic approximations.

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

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