Lecture 10: Hypothesis Testing

Ma Xuejun

School of Mathematical Sciences

Soochow University

https://xuejunma.github.io

Outline

1. Hypothesis Testing
2. The Neyman-Pearson Test
3. The Wald Test
4. The Likelihood Ratio Test (LRT)
5. Three tests
6. p-values
7. The Permutation Test
8. Multiple Testing Problem: FWE, FDR, HC.

Hypothesis Testing

- We do not need good estimation of the parameter; we are interested in one value only
- To test the effects of two medicine, we are interested in the difference of the effect equals to 0 or not

Formalize it and we state it as a null hypothesis H_0 and an alternative hypothesis H_1. For example,

$$H_0: \ \theta = \theta_0 \quad versus \quad H_1: \ \theta \neq \theta_0$$

Generally,we want to test

$$H_0: \ \theta \in \Theta_0 \quad versus \quad H_1: \ \theta \in \Theta_1$$

Where $\Theta_0 \cap \Theta_1 = \emptyset$. If $\Theta_0 = \{\theta\}$, it is called a *simple null hypothesis*, otherwise, it is a *composite null hypothesis*

Hypothesis Testing

For a hypothesis testing problem:

- Underlying truth: H_0 is true or H_1 is true
- Goal: sufficient evidence to reject H_0?
- Action: reject H_0 or not reject H_0

	Decision	
	Retain H_0	Retain H_0
H_0 is true	✓	Type I error(false positive)
H_1 is true	Type II error(false positive)	✓

- Without sufficient evidence, we do not reject H_0. It does not mean we believe it is correct
- Obviously, the setting prefers H_0

Hypothesis Testing

Example. $X_1, \cdots, X_n \sim Bernoulli(p)$. Then the problem is

$$H_0: \ p = 1/2 \quad versus \quad H_1: \ p \neq 1/2.$$

- What is a test?
- A test need a statistic T and a rejection region R. If $T \in R$ then we reject H_1.
- For example, let $T = \bar{X}$ and the rejection be $(0, 0.3) \cup (0.6, 1)$, then the test is

$$Reject\ H_0\ if\ |\bar{X} - 1/2| > 0.1.$$

- With the data, we can claim whether we reject H_0 or not
- With this test, Type I error is $P(|\bar{X} - 1/2| > 0.1|H_0)$, Type II error is $1 - P(|\bar{X} - 1/2| > 0.1|H_1)$
- There are multiple tests for one hypothesis testing problem

Evaluation of a test

- With a test, we hope we can do correct justifications.
- It means minimizing the Type I error and Type II error.
- For the type II error, we define the power function.

Definition: Power function

Suppose we reject H_0 when $T(X_1, \cdots, X_n) \in R$. The *power function* is defined as

$$\beta(\theta) = P(T(X_1, \cdots, X_n) \in R|\theta).$$

Remark.

- The power function is a function about θ
- When $\theta \in \Theta_1$, it measures the probability that the test correctly rejects H_0
- When $\theta \in \Theta_0$, it measures the type I error.

Evaluation of a test

- For the type I error, one way is to control it with the maximum

Definition

A test is *size α* if

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

A test is *level α* if

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

- A *size* α test and a *level* α test are almost the same thing. The distinction is made because sometimes we want a *size* α test and we cannot construct a test with exact *size* α. But we can build one with smaller error rate
- Motivation: Type I error is not the same important with the Type II error. say, for medical diagnosis, we should minimize the Type II error(discover people with disease correctly), and control Type I error (healthy people are labeled with disease) at a low level.
- Common values for $\alpha: 0.01, 0.05, 0.1$

Evaluation

The general strategy to construct a test is

(1) Fixe $\alpha \in [0, 1]$

(2) Try to maimize $\beta(\theta)$ for $\theta \in \Theta_0$, subject to $\beta(\theta) \leq \alpha$ for $\theta \in \Theta_0$

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

$$H_0 : \theta = \theta_0 \quad versus \quad H_1 : \theta > \theta_0.$$

This is called a $\underline{one - sided\ alternative}$. Suppose we reject H_0 if $T_n > c$ where

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

Then, the power function is

$$\beta(\theta) = P\Big(\frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}} > c; \theta\Big) = P\Big(\frac{\bar{X}_n - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}; \theta\Big)$$
$$= P\Big(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\Big) = 1 - \Phi(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\Big),$$

where $Z \sim N(0, 1)$ and Φ is the CDF for Z. Now,

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

Evaluation

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

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

Our test is to reject H_0 when $T_n = \frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}} > z_\alpha$.
Now, let's consider the two-sided alternative, that

$$H_0: \ \theta = \theta_0 \quad versus \quad H_1: \ \theta \neq \theta_0.$$

We will reject H_0 if $|T_n| > c$. The power function is

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

The size is $\beta(\theta_0) = 2\Phi(-c)$. Let it equal to α,then
$c = -\Phi^{-1}(1 - \alpha/2) = z_{\alpha/2}$. The test is to reject H_0 when $|\frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}}| > z_{\alpha/2}$

Generally used tests

There are some tests that are found to be useful or optimal:

- Neyman-Pearson Test
- Wald Test
- Likelihood Ratio Test (LRT)
- Score Test
- Permutation Test
- Bootstrap Test

Now we discuss them one by one.

The Neyman-Pearson Test

- The Neyman-Pearson test considers only simple null and simple alternative, which means the test
$$H_0: \ \theta = \theta_0 \quad versus \quad H_1: \ \theta = \theta_1.$$

Definition: Neyman-Pearson Test

Let $L(\theta) = f(X_1, \cdots, X_n; \theta)$ and
$$T_n = L(\theta_1)/L(\theta_0).$$
Suppose we reject H_0 if $T_n > k$ where k is chosen so that
$$P\Big(T(X_1, \cdots, X_n) > k; \theta = \theta_0\Big) = \alpha,$$
then it is called a Neyman-Pearson Test.

- The test statistic is the ratio of two joint densities. It is to check with which likelihood, the data is more possible.
- It is quite limited, since it requires both the null and the alternative are simple.

The Neyman-Pearson Test

Definition: Uniformly Powerful Tests

Let C_α be a collection of level α for $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$. A test in C_α with power function $\beta(\theta)$ is uniformly most powerful (UMP) if for every $\beta'(\theta)$ which is the power function of any other test in C_α, then

$$\beta(\theta) \geq \beta(\theta'), \qquad \theta \in \Theta_1.$$

- It is possible that a UMP does not exist
- In the simple null and simple alternative case, it exists, which is the Neyman-Pearson test.

Neyman-Pearson Lemma

Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Then

- The Neyman-Pearson test is a UMP level α test;
- If such a test exists, then every UMP level α test is a Neyman-Pearson test.

Example

Example. Let $X_1 \sim Bernomial(2, \theta)$ and we want to test

$$H_0: \ \theta = 1/2 \quad versus \quad H_1: \ \theta = 3/4.$$

We have that

$$\frac{f(0; 3/4)}{f(0; 1/2)} = \frac{1}{4}; \quad \frac{f(1; 3/4)}{f(1; 1/2)} = \frac{3}{4}; \quad \frac{f(2; 3/4)}{f(2; 1/2)} = \frac{9}{4}$$

If we construct a test that reject H_0 when

$$\frac{f(X_1; 3/4)}{f(X_1; 1/2)} > 2$$

Then the test has level as

$$P(\frac{f(X_1; 3/4)}{f(X_1; 1/2)} > 2; \theta_0) = P(X_1 = 2; \theta_0) = 1/4.$$

So it is a UMP level $1/4$ test.

The Wald Test

- Assume there is an asymptotic normal estimator $\hat{\theta}_n$, where $\hat{\theta}_n - \theta \xrightarrow{d} N(0, \sigma_n^2)$
- If $H_0: \ \theta = \theta_0$ is true, then there is $\quad \hat{\theta}_n - \theta \xrightarrow{d} N(0, \sigma_n^2)$
- we can construct a test statistic

$$T_n = \frac{\hat{\theta}_n - \theta_0}{\hat{\theta}_n}$$

- If H_0 is true, $T_n \xrightarrow{d} N(0,1)$, which concentrates at 0. So, if T_n is too large/small, we reject H_0.
- This kind of test is called the Wald Test.

Example.

- With Bernoulli data, to test $H_0: \ p = p_0$ and $H_1: \ p \neq p_0$, recall that $\sqrt{n}(\bar{X} - p) \xrightarrow{d} N(o, p(1-p))$, we can construct a Wald test

$$T_n = |\frac{\bar{X} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}}| > c,$$

where $c = \Phi^{-1}(1 - \alpha/2) = z_{\alpha/2}$.

The Wald Test

- Consider MLE $\hat{\theta}_n$. According to the asymptotic normality of MLE, there is
$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{1/I(\theta)}} \xrightarrow{d} N(0,1).$$
So we can construct a test w.r.t. MLE, which is to reject null hypothesis when
$$T_n = \frac{\hat{\theta}_n - \theta_0}{\sqrt{1/nI(\hat{\theta})}} > c.$$
- If it happens that $\bar{X}$ is an estimator for θ. According to CLT, $\sqrt{n}(\bar{X} - \theta)/\sigma \xrightarrow{d} N(0,1)$. So we can also build a Wald test based on the average
- Usually, σ_n is a function of θ. Since the truth is unknown, we can either apply θ_0 or $\hat{\theta}$ in practice.
- The Wald test requires asymptotic normality, so it works for $large$ sample size only.

The Likelihood Ratio Test

- Neymann-Pearson test is the ratio of likelihoods w.r.t. two values
- For composite null and alternative, we can generalize the idea

Definition: Likelihood Ratio Test (LRT)

The LRT statistic for testing $H_0 : \theta \in \Theta_0 \quad versus \quad H_1 : \theta \in \Theta_1$ is

$$\lambda(x_1, \cdots, x_n) = \frac{\sup_{\theta \in \Theta_0} f_{X_{1:n}}(x_1, \cdots, x_n; \theta)}{\sup_{\theta \in \{\Theta = \Theta_0 \bigcup \Theta_1\}} f_{X_{1:n}}(x_1, \cdots, x_n; \theta)}$$

A LRT is any test that has a rejection region of the form $\{(x_1, \cdots, x_n); \lambda(x_1, \cdots, x_n) \leq c\}$ for any constant $c \in [0, 1]$.

- Θ_0: null parameter space; Θ: the whole parameter space
- According to the definition of MLE, the LRT statistic can be written as

$$\lambda(x_1, \cdots, x_n) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

- If it is small, then we reject H_0.

LRT: Example

Example. Suppose that $X_i \overset{i.i.d}{\sim} N(\theta, 1)$ and suppose we want to test
$H_0: \ \theta = \theta_0 \quad versus \quad H_1: \ \theta \neq \theta_0$.
Recall that the MLE is $\hat{\theta} = \bar{X}_n$. So the LRT statistic is

$$\lambda(x_1, \cdots, x_n) = \frac{\frac{1}{(\sqrt{2\pi})^n} e^{-\frac{\sum (x_i - \theta_0)^2}{2}}}{\frac{1}{(\sqrt{2\pi})^n} e^{-\frac{\sum (x_i - \bar{x})^2}{2}}} = \frac{\exp\{-\frac{\sum (x_i - \theta_0)^2}{2}\}}{\exp\{-\frac{\sum (x_i - \bar{x})^2}{2}\}}.$$

Since $\sum (x_i - \theta_0)^2 = \sum (x_i - \bar{x})^2 + n \sum (\bar{x} - \theta_0)^2$, we have
$\lambda(x_1, \cdots, x_n) = \exp\{-\frac{n}{2}(\bar{x} - \theta)^2\}$. Since it is monotone with $|\bar{x} - \theta_0|$,so the rejection region is equivalent with

$$\{x \in \mathbb{R}^n : |\bar{x} - \theta_0| \geq c\}.$$

Since $\bar{x} - \theta_0 \sim N(0, 1/n)$ under null hypothesis, the level of the test is $2\Phi(-\sqrt{n}c)$. For a level α test, we have $c = \Phi^{-1}(1 - \alpha/2)/\sqrt{n}$.

- **Example 8.2.3 (Exponential LRT)** $X_1, X_2, \cdots, X_n$ be a random sample from $f(x|\theta) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta, \\ 0, & x < \theta. \end{cases}$

$$L(\theta|\mathbf{x}) = \begin{cases} e^{-\sum x_i + n\theta}, & \theta \leq x_{(1)}, \\ 0, & \theta > x_{(1)}. \end{cases}$$

$H_0 : \theta \leq \theta_0$, versus $H_1 : \theta > \theta_0$

$$\sup_{\theta \leq \theta_0} L(\theta|\mathbf{x}) = \begin{cases} e^{-\sum x_i + n\theta_0}, & \theta < x_{(1)}, \\ e^{-\sum x_i + n x_{(1)}}, & \theta \geq x_{(1)}. \end{cases}$$

$$\sup_{\theta} L(\theta|\mathbf{x}) = e^{-\sum x_i + n x_{(1)}}.$$

Therefore,

$$\begin{aligned}\lambda(\mathbf{x}) &= \begin{cases}1, & x_{(1)} \le \theta_0, \\ e^{-n(x_{(1)}-\theta_0)}, & x_{(1)} > \theta_0.\end{cases} \\ \{\mathbf{x} : \lambda(\mathbf{x}) \le c\} &= \{\mathbf{x} : x_{(1)} \ge \theta_0 - \frac{\log c}{n}\}\end{aligned}$$

Note that the rejection region depends on the sample only through $x_{(1)}$.

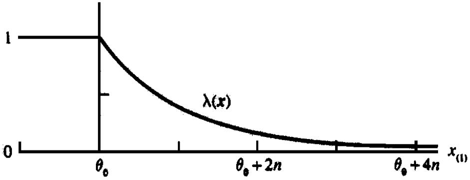


Figure 8.2.1 $\lambda(x)$: *a function only of* $x_{(1)}$

LRT: Theorem

- Can we always find a proper distribution for the LRT statistic?
- Not always, but asymptotically, yes.

Theorem: LRT statistics

Let $X_i \overset{i.i.d}{\sim} F_X(\cdot;\theta^*)$ with $f_X(\cdot;\theta^*)$ as the associated PDF. Let $\hat{\theta}_n$ be the MLE. Consider the testing $H_0 : \theta = \theta_0 \quad versus \quad H_1 : \theta \neq \theta_0$ where $\theta \in \mathbb{R}$.
The under H_0,

$$-2\log(\lambda(X_1, X_2, \cdots, X_n)) \xrightarrow{d} {\chi_1}^2.$$

- Regularity conditions for MLE normality
- To construct a level α test, we can set the rejection region as $-2\log(\lambda(X_1, X_2, \cdots, X_n)) \geq \chi^2_{1,\alpha}$, where ${\chi^2}_{1,\alpha}$ is the $1-\alpha$ quantile for ${\chi^2}_{1,\alpha}$ distribution.
- If $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$. Then, under the regularity conditions,

$$T_n = -2\log(\lambda(X_1, X_2, \cdots, X_n)) \xrightarrow{d} {\chi^2}_v, \quad v = \dim(\Theta) - \dim(\Theta_0).$$

Proof

- Under the regularity conditions, we have the Taylor expansion for the log-likelihood function $l(\theta)$ close to the point $\hat{\theta}$:

$$l(\theta) \approx l(\hat{\theta}) + l'(\hat{\theta})(\theta - \hat{\theta}) + l''(\hat{\theta})\frac{(\theta - \hat{\theta})^2}{2} = l(\hat{\theta}) + l''(\hat{\theta})\frac{(\theta - \hat{\theta})^2}{2}$$

- The expression for LRT statistic is

$$\begin{aligned} -2\log(\lambda(X_1, X_2, \cdots, X_n)) &= -2l(X_1, X_2, \cdots, X_n; \theta_0) \\ &\quad + 2l(X_1, X_2, \cdots, X_n; \hat{\theta}) \\ &\approx 2l(\hat{\theta}) - 2l(\hat{\theta}) - l''(\hat{\theta})(\theta - \hat{\theta})^2 \\ &= -l''(\hat{\theta})(\theta - \hat{\theta})^2 \\ &= \frac{-l''(\hat{\theta})}{I_n(\theta_0)} \times I_n(\theta_0)(\hat{\theta} - \theta_0)^2 = A_n \times B_n \end{aligned}$$

- Note that $A_n \xrightarrow{P} 1$ according to WLLN and $\sqrt{B_n} \xrightarrow{d} N(0,1)$, so that $B_n \xrightarrow{d} {\chi_1}^2$. According to Slutsky's theorem, the result follows.

LRT: Examples

- $X_i, \cdots, X_n \overset{i.i.d}{\sim} Poission(\lambda)$.
- The log-likehood function is $\sum X_i \log \lambda - n\lambda + C$.
- We want to test $H_0 : \ \lambda = \lambda_0 \quad versus \quad H_1 : \ \lambda \neq \lambda_0$.
- Recall the MLE is $\hat{\lambda} = \sum X_i / n$, then the LRT statistic is

$$-2\log(\lambda(X_1, X_2, \cdots, X_n)) = 2n[(\lambda_0 - \hat{\lambda}) - \hat{\lambda}\log(\lambda_0/\hat{\lambda})]$$

- we reject H_0 when $-2\log(\lambda(X_1, X_2, \cdots, X_n)) > {\chi^2}_{1,\alpha}$.

LRT: Examples

- Consider a multinomial distribution with $\theta = (p_1, p_2, \cdots, p_5)$, So

$$L(\theta) = p^{y1} \cdots p^{y5}, \qquad y_k = \sum 1\{X_i = k\}, k = 1, 2, 3, 4, 5.$$

Suppose we want to test

$$H_0 : \ p_1 = p_2 = p_3 \ and \ p_4 = p_5 \quad versus \quad H_1 : \ H_0 \ is \ false.$$

Then $v = \dim(\Theta) - \dim(\Theta_0) = 4 - 1 = 3$. The LRT test statistic is

$$\lambda(x_1, \cdots, x_n) = \frac{\prod_{i=1}^{5} \hat{p_0}_j^{Y_j}}{\prod_{i=1}^{5} \hat{p}_j^{Y_j}}$$

where $\hat{p}_j = Y_j/n, \hat{p}_{10} = \hat{p}_{20} = \hat{p}_{30} = (Y_1 + Y_2 + Y_3)/n$, $\hat{p}_{40} = \hat{p}_{50} = (1 - 3\hat{p}_{10})/2$. We reject H_0 if the test statistic is larger than ${\chi^2}_{3,\alpha}$

Three tests

We start with the simplest case of iid data with one unknown real parameter. Then for testing

$$H_0 : \theta = \theta_0 \quad H_a : \theta \neq \theta_0$$

- Wald test

$$T_W = \frac{(\hat{\theta}_{MLE} - \theta_0)^2}{\left[I_T(\hat{\theta}_{MLE})\right]^{-1}} = (\hat{\theta}_{MLE} - \theta_0)^\top I_T(\hat{\theta}_{MLE})(\hat{\theta}_{MLE} - \theta_0)$$

- Likelihood ratio test

$$T_{LR} = -2 \log \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})} = -2\left[\ell(\theta_0) - \ell(\hat{\theta}_{MLE})\right]$$

- Score test

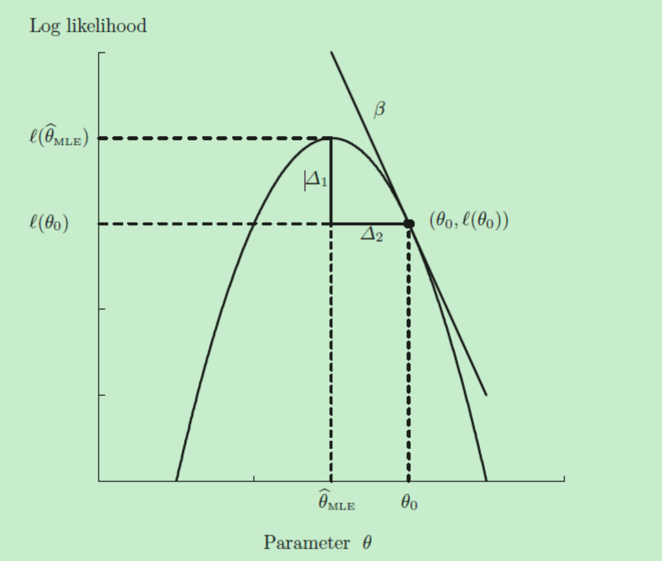
$$T_S = \frac{S^2(\theta_0)}{I_T(\theta_0)} = S^\top(\theta_0)[I_T(\theta_0)]^{-1} S(\theta_0)$$

Note: $S(\theta) = \frac{\partial}{\partial \theta^\top} \ell(\theta)$, $I_T(Y, \theta) = \frac{\partial}{\partial \theta} S(\theta)$, $I_T(\theta) = E[I_T(Y, \theta)] = nI(\theta)$.

- $$[I_T(\theta_0)]^{-1/2}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, \mathbb{I}_p)$$
- Under H_0, $I_T(\hat{\theta}_{MLE})[I_T(\theta_0)]^{-1} \xrightarrow{p} \mathbb{I}_p$. Hence
$$T_W = (\hat{\theta}_{MLE} - \theta_0)^\top I_T(\hat{\theta}_{MLE})(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} \chi^2(p)$$
- Under H_0, $S(\theta_0)$ has mean 0, variance $I_T(\theta_0)$. Hence $[I_T(\theta_0)]^{-1/2}S(\theta_0) \xrightarrow{p} \mathbb{I}_p$, and
$$T_S = S^\top(\theta_0)[I_T(\theta_0)]^{-1}S(\theta_0) \xrightarrow{d} \chi^2(p)$$
- $$\ell(\theta_0) = \ell(\hat{\theta}_{ML}) + S(\hat{\theta}_{ML}) - \sqrt{n}(\hat{\theta}_{MLE} - \theta_0)^\top \frac{1}{2} I_n(Y, \hat{\theta}^*)\sqrt{n}(\hat{\theta}_{MLE} - \theta_0)$$
where $\hat{\theta}^*$ lies between $\hat{\theta}_{ML}$ and θ_0. $I_n(Y, \hat{\theta}^*) \xrightarrow{p} I(\theta_0)$

1

[1] Essential Statistical inference Theory and Methods, Dennis D. Boos and L. A. Stefanski


Log likelihood
β
$\ell(\widehat{\theta}_{\mathrm{MLE}})$
Δ_1
$\ell(\theta_0)$
$(\theta_0, \ell(\theta_0))$
Δ_2
$\widehat{\theta}_{\mathrm{MLE}}$
θ_0
Parameter θ

Normal model with known variance

Suppose that $Y_1, \ldots, Y_n$ iid $N(\mu, 1)$. $H_0 : \mu = \mu_0$, then

$$\ell(\mu) = \log L(\mu|Y) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^2$$

$$S(\mu) = \frac{\partial}{\partial \mu}\ell(\mu) = \sum_{i=1}^{n}(Y_i - \mu)$$

$$I_T(Y, \mu) = \frac{\partial}{\partial \mu}S(\mu) = n$$

So that $M\hat{L}E = \overline{Y}$, and $I_T(\mu) = E[I_T(Y, \mu)] = n$. Hence

$$T_W = \frac{(\overline{Y} - \mu_0)^2}{n^{-1}} = (\overline{Y} - \mu_0)^2$$

$$T_S = \frac{\left[\sum_{i=1}^{n}(Y_i - \mu)\right]^2}{n} = (\overline{Y} - \mu_0)^2$$

$$T_{LR} = -2\Big[-\frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu_0)^2 + -\frac{1}{2}\sum_{i=1}^{n}(Y_i - \bar(Y))^2\Big] = (\overline{Y} - \mu_0)^2$$

p-values

- Given α, we construct a level α test
- With data, we calculate the statistic and decide whether to $reject$ or $retain\ H_0$
- If α changes, should we do all the steps again?

Definition: P-values

A p-value $p(X)$ is a test statistic with $p(X) \in [0,1]$. Small values of p indicate that H_1 is true. A p-value is valiid if for every $\theta \in \Theta_0, \alpha \in [0,1]$,

$$P(p(X) \leq \alpha; \theta) \leq \alpha.$$

- $p(X)$ is a test statistic. With the statistic $p(X)$, the level α test is to reject H_0 when $p < \alpha$. The power function w.r.t. this test is

$$\beta(\theta) = P(p(X) \leq \alpha; \theta).$$

- Therefore, it can be **viewed as the smallest α at which we would reject H_0**.

p-values

- Question: how to find this test statistic?

Theorem: P-values

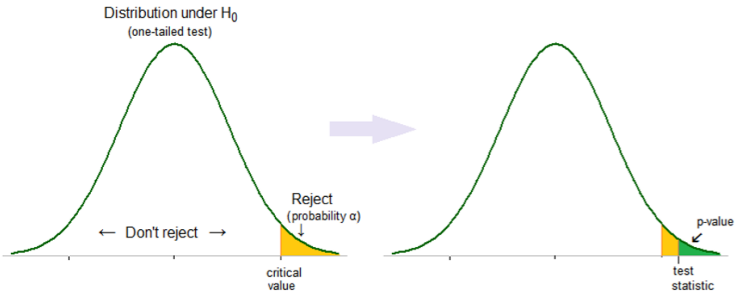
Let $W(X)$ be a test statistic such that large values of W indicate that H_1 is true. For each $x \in X$, define

$$p(x) = \sup_{\theta \in \Theta_0} P(W(X) \geq W(x); \theta),$$

then $p(X)$ is a valid p-value.

- Note that $W(x)$ should satisfy that reject H_0 when $T(x) > c$.
- This is the general way to find the p-value. We define a test statistic first, and then define p be the probability that the statistic is no smaller than the observation.
- p-value may change for different test statistic, even with the same data. So when we specify p-value, we should specify the test statistic.

Remarks



- p-value is the probability for the test statistic under null, not the probability of H_0
- Why p-value is useful, not the test statistic?

Theorem

Under $H_0, p \sim Unif(0,1)$

We never know what the test statistic means, but we can achieve the information in p-value quickly

Example: p-values

Example. Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(0,1)$. Test that $H_0: \ \theta = \theta_0 \quad versus \quad H_1: \ \theta \neq \theta_0$. We reject when $|T_n| = |\sqrt{n}(\bar{X}_n - \theta_0)|$ is large. Let t_n be the observed value of T_n. Let $Z \sim N(0,1)$. Then,

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

Now, we can return the p-value to the researcher, with which the researcher can easily tell how strong the evidence is to reject H_0.

The Permutation Test

Suppose we have $X_1, \cdots, X_n \sim F$ and $Y_1, \cdots, Y_m \sim G$. We want to test

$$H_0: \ F = G \quad versus \quad H_1: \ F \neq G$$

(1) Let $Z = (X_1, \cdots, X_n, Y_1, \cdots, Y_m)$. Greate labels as $L = (1, 1, \cdots, 1, 1, \cdots, 2)$, where there are n 1's and m 2's. So L are the label for the observation.

(2) The test statistic can be written as a function of Z and L. For example, $|\bar{X}_n - \bar{Y}_m|$ can be written as

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

So $T = g(L, Z)$.

(3) Define: $p = \frac{1}{(n+m)!} \sum_{\pi} I(g(L_\pi, Z) > g(L, Z))$, where L_π is a permutation of the labels and the sum is over all permutations.

(4) Under $H_0, F = G$, so the distribution of T does not change, and $p \sim Unif(0, 1)$ (discrete version).

(5) Reject H_0 if $p < \alpha$.

The Permutation Test

Summing over all permutations is infeasible for large data sets. The computation load is $(n+m)!$. Usually, it suffices to use a random sample of the permutations. So the procedure becomes

(1) Let $Z = (X_1, \cdots, X_n, Y_1, \cdots, Y_m)$. Create labels as $L = (1, 1, \cdots, 1, 1, \cdots, 2)$, where there are n 1's and m 2's.

(2) Let $T = g(L, Z)$. Compute a random permutation of the labels $\pi_i, i = 1, \cdots, K$. Define

$$p = \frac{1}{K} \sum_{i=1}^{K} I(g(L_{\pi_i}, Z) > g(L, Z))$$

(3) Reject H_0 if $p < \alpha$.

- Distribution free
- Does not involve any asymptotic approximation
- Flexible to derive for any statistics

Tukey's story

John Tukey is a famous mathematician and statistician, well known for the development of FFT algorithm and box-plot. When he taught in Princeton University, a young scientist came to him and asked him one question.

- Scientist: I administers 250 uncorrelated tests, where 11 were significant at the 5% level. Should I claim that these 11 were really significant?
- Tukey: No, we expect

$$250 \times 5\% = 12.5 \quad significances\ at\ the\ 5\%\ level$$

- Now the question comes: when we can claim significance if we have such a problem? The significance level α does not work?

Multiple Testing

- The classical hypothesis testing problem is to consider limited parameters, say $\theta = \theta_0$ or no
- Sometimes, we need to do multiple testing at the same time, say

$$\begin{aligned} H_{10}: \ \theta_1 = 0 \quad versus \quad H_{11}: \ \theta_1 \neq 0 \\ H_{20}: \ \theta_2 = 0 \quad versus \quad H_{21}: \ \theta_2 \neq 0 \\ \ldots \quad \ldots \quad \ldots \\ H_{N0}: \ \theta_N = 0 \quad versus \quad H_{N1}: \ \theta_N \neq 0 \end{aligned}$$

- For example, we want to identify the genes that cause one specific disease. For each gene, we want to know whether it works or not. The number of genes is pretty large here.
- Can we do the hypothesis testing problems individually, and reject H_0 if any individual H_0 is rejected?
- No and Yes.

Multiple Testing

- Recall. For a level α test, we have $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.
- Consider individual testing problem:

$$H_{i0}: \ \theta_i = 0 \quad versus \quad H_{i1}: \ \theta_i \neq 0$$

  Suppose we have a level α test for this problem. Denote the power function of this test as $\beta_i(\theta)$. Then $\beta_i(\theta) \leq \alpha$.
- Therefore, for the original multiple testing problem, the rejection probability is

$$\beta(\theta) = 1 - P(accept\ H_{i0}\ for\ all\ i \leq i \leq N) \overset{indep}{=} 1 - \prod_{i=1}^{N} (1 - \beta_i(\theta))$$

- Under null hypothesis $\theta = 0$, if all the test statistic are independent and all the tests have size α. The rejection probability is

$$1 - \prod_{i=1}^{N} (1 - \beta_i(0)) = 1 - (1 - \alpha)^N$$

  How large it is? Let $N = 50, \alpha = 0.05$, then $1 - (1 - \alpha)^N \approx 0.92$. $P(rejection | H_0) = 0.92$!

Familywise Error Control

We need to adjust the level for individual tests

- Define $I = \{i; H_{i0}\ is\ true\}$ be the index set for which H_0 is true
- Define $R = \{i; H_{j0}\ is\ rejected\}$ be the index set that we reject.
- We say that we have controlled the familywise error rate at level α if

$$P(R \cap I \neq \emptyset) = P((making\ a\ false\ rejection) \leq \alpha.$$

- **Bonferroni method**:for each individual test, set the level to be α/N. Let p_j be the p-value for test H_{j0} versus H_{j1}.

$$\begin{aligned} P(making\ a\ false\ rejection) &= P(p_j < \alpha/N\ for\ some\ i \in I) \\ &\leq \sum_{i \in I} P(p_j < \alpha/N) \\ &= \sum_{i \in I} \alpha/N = \frac{\alpha|I|}{N} \leq \alpha \end{aligned}$$

  So we have overall control of the type I error.
- It can have low power.

Normal Example

Example. Suppose we have N sample means $Y_1, \cdots, Y_N$, each is the average of n normal observations with variance σ^2. So $Y_j \sim N(\mu_j, \sigma^2/n)$. To test $H_{j0}: \ \mu_j = 0$ we can use the test statistic

$$T_j = \sqrt{n}Y_j/\sigma \sim N(\mu_j, 1).$$

The power function at $\mu_j = 0(p - value)$ is $p_j = 2\Phi(-|T_j|)$.

- If we did uncorrected testing that we reject $p_j < \alpha$, which means $|T_j| > z_{\alpha/2}$.
- With Bonferroni correction we reject when $p_j < \alpha/N$, which corresponds to

$$|T_j| > z_{\alpha/2N}$$

- Generate random samples under H_0 with code in next slide.
- If we apply the approximation for normal CDF and PDF, tha

$$\frac{\phi(x)}{x + 1/x} \leq 1 - \Phi(x) \leq \frac{\phi(x)}{x}, \ \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

then approximately, the corrected bound becomes $\sigma\sqrt{2\log(2N/\alpha)/n}$. It grows like $\sqrt{\log N}$.

```
N = 50; n = 100; sigma = 3; alpha = 0.05;
iter = 100; fwr1 = rep(0, iter); fwr2 = rep(0, iter);
for(i in 1:iter){
  Y = rnorm(N, mean = 0, sd = sigma/sqrt(n));
  #Generate Y_i's under null
  stat = sqrt(n)*Y/sigma;
  #Calculate the test statistic T_i
  p = 1*(abs(stat) > qnorm(1 - alpha/2))
  #Find the p-value for each individual test without correction
  corp = 1*(abs(stat) > qnorm(1 - alpha/2/N))
  #Find the p-value for each individual test with Bonferroni correct
  fwr1[i] = 1*(sum(p) > 0); #Familywise error for test 1;
  fwr2[i] = 1*(sum(corp) > 0); #Familywise error for test 2;
}
mean(fwr1) #empirical familywise error for uncorrected test
mean(fwr2) #empirical familywise error for corrected test
```

Thank you!