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. Bootstrap Resampling
   - Permutation and Rank Tests
8. Multiple Testing Problem: FWE, FDR, HC.
   - Familywise Error Control
   - Higher Criticism

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 $size\ \alpha$ test and a $level\ \alpha$ test are almost the same thing. The distinction is made because sometimes we want a $size\ \alpha$ test and we cannot construct a test with exact $size\ \alpha$. 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 $one - sided\ alternative$. 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 $size$ α 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 8.3.11 :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

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

- The requirements in Definition 8.3.11 are so strong that UMP tests may not exist in realistic problem.
- In the simple null and simple alternative case, it exists, which is the Neyman-Pearson test.

- **Theorem 8.3.12 (Neyman-Pearson Lemma)** Consider testing $H_0 : \theta = \theta_0$ v.s. $H_1 : \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(\mathbf{x}|\theta_i)$, $i = 0, 1$, using a test with rejection region R that satisfies

$$\mathbf{x} \in R \text{ if } f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0) \tag{2.1}$$

and

$$\mathbf{x} \in R^c \text{ if } f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0) \tag{2.2}$$

for some $k \geq 0$ and

$$\alpha = P_{\theta_0}(\mathbf{X} \in \mathbf{R})$$

Then

a. (*Sufficiency*) Any test that satisfies (2.1) and (2.2) is a UMP level α test,
b. (*Necessity*) If there exists a test that satisfying (2.1) and (2.2) with $k > 0$, then every UMP level α test is a size α test (satisfies (2.2)) and every UMP level α test satisfies (2.2) except perhaps on a set A satisfying $P_{\theta_0}(\mathbf{X} \in \mathbf{A}) = \mathbf{P}_{\theta_1}(\mathbf{X} \in \mathbf{A}) = \mathbf{0}$.

- **Corollary 8.3.13** Suppose that $T(\mathbf{x})$ is a sufficient statistic for θ and $g(t|\theta_i)$ is the pdf or pmf of T corresponding to θ_i, $i = 0, 1$. Then any test based on T with rejection region S (a subset of the sample space of T) is a UMP level α test if it satisfies

$$t \in S \text{ if } g(t|\theta_1) > kg(t|\theta_0)$$

and

$$t \in S^c \text{ if } g(t|\theta_1) < kg(t|\theta_0)$$

for some k, where

$$\alpha = P_{\theta_0}(T \in S) \tag{2.3}$$

Example 8.3.14 (UMP binomial test) Let $X \sim \text{Binomial}(2, \theta)$. We want to test $H_0 : \theta = \frac{1}{2}$, versus $H_1 : \theta = \frac{3}{4}$. Calculating the ratios of the pmfs gives

$$\frac{f(0|\theta = \frac{1}{2})}{f(0|\theta = \frac{3}{4})} = \frac{1}{4}, \quad \frac{f(1|\theta = \frac{1}{2})}{f(1|\theta = \frac{3}{4})} = \frac{3}{4}, \text{ and } \frac{f(2|\theta = \frac{1}{2})}{f(2|\theta = \frac{3}{4})} = \frac{9}{4}.$$

- $\frac{3}{4} < k < \frac{9}{4}$, the Neyman-Pearson Lemma says that the test that rejects H_0 if $X = 2$ is the UMP level $\alpha = P(X = 1|\theta = \frac{1}{2}) = \frac{1}{4}$ test.
- $\frac{1}{4} < k < \frac{3}{4}$, the Neyman-Pearson Lemma says that the test that rejects H_0 if X=1 or 2 is the UMP level $\alpha = P(X = 1 \text{ or } 2|\theta = \frac{1}{2}) = \frac{3}{4}$ test.
  ...

- **Example 8.3.15 (UMP normal test)** Let $X_1, \cdots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population, σ^2 known.
- The sample mean $\bar{X}$ is a sufficient statistic for θ.
- Consider testing $H_0 : \theta = \theta_0$, versus $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$.
- The inequality $g(\bar{x}|\theta_1) > kg(\bar{x}|\theta_0)$, is equivalent to

$$\bar{x} < \frac{(2\sigma^2 \log k)/n - \theta_0^2 + \theta_1^2}{2(\theta_0 - \theta_1)}.$$

  The fact that $\theta_1 - \theta_0 < 0$ was used to obtain the inequality.
  - The right-hand side increases from $-\infty$ to ∞ as k increases from 0 to ∞.
  - Thus, by Corollary 8.3.13, the test with rejection region $\bar{x} < c$ is the UMP level α test, where $\alpha = P_{\theta_0}(\bar{X} < c)$.
  - If a particular α is specified, then the UMP test rejects H_0 if $\bar{X} < c = -\sigma z_\alpha/\sqrt{n} + \theta_0$. This choice of c ensures that (2.3) is true.

- In the case of composite hypothesis, the UMP test can be derived with the Neyman-Pearson Lemma.

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 $\hat{\theta}_n - \theta \xrightarrow{d} N(0, \sigma_n^2)$
- 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 $\{(x_1, \cdots, x_n); \lambda(x_1, \cdots, x_n) \le c\}$ 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,

$$\lambda(\mathbf{x}) = \begin{cases} 1, & x_{(1)} \leq \theta_0, \\ e^{-n(x_{(1)}-\theta_0)}, & x_{(1)} > \theta_0. \end{cases}$$

$$\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\} = \{\mathbf{x} : x_{(1)} \geq \theta_0 - \frac{\log c}{n}\}$$

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

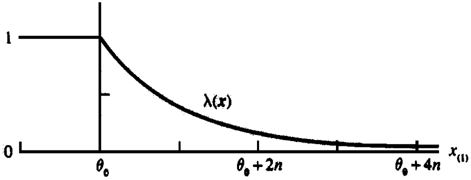


Figure 8.2.1 $\lambda(x)$: *a function only of* $x_{(1)}$

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.

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

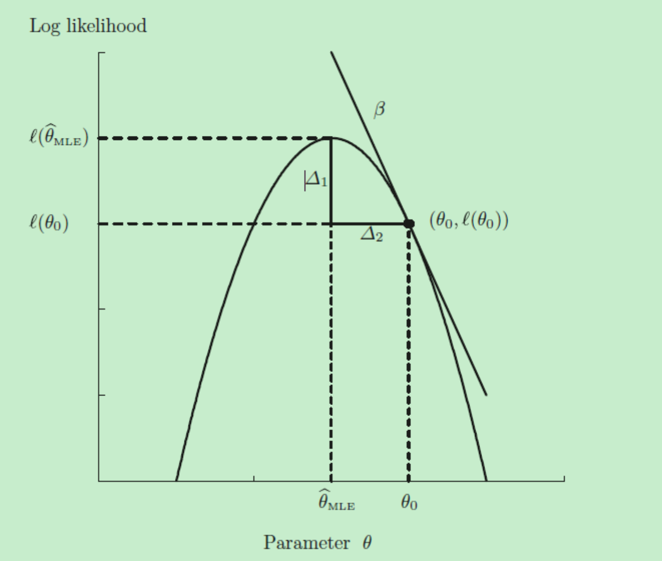
$$T_S = \frac{S^2(\theta_0)}{I_T(\theta_0)} = S^\top(\theta_0)[I_T(\theta_0)]^{-1} S(\theta_0)$$

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
$\ell(\widehat{\theta}_{\rm MLE})$
$\ell(\theta_0)$
β
Δ_1
Δ_2
$(\theta_0, \ell(\theta_0))$
$\widehat{\theta}_{\rm MLE}$
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

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

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

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

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

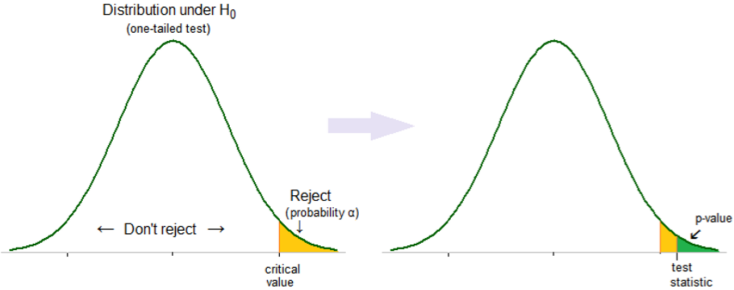
Let $W(X)$ be a test statistic such that large values of W indicate that H_1 is true. For each $x \in X$, define

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

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

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.

Bootstrap Resampling

Example 11.4 (Comparing two variances).

- We again consider the two independent samples situation: $X_1, ..., X_m$ and $Y_1, ..., Y_n$ with respective means μ_X and μ_Y and variances σ_X^2 and σ_Y^2.
- The null hypothesis of interest is $H_0 : \sigma_X = \sigma_Y$.
- A semiparametric assumption is made that both samples are from the same location-scale family, but the family is unknown.
- Thus, the distribution function of each X_i is $F_1(x) = F_0((x - \mu_X)/\sigma_X)$ and the distribution function of each Y_i is $F_2(x) = F_0((x - \mu_Y)/\sigma_Y)$, where F_0 is the cdf of an unknown distribution with mean 0 and variance 1.
- The statistic used to test H_0 is

$$T_{mn} = \left(\frac{mn}{m+n}\right)^{1/2} \{\log(s_X^2) - \log(s_Y^2)\}.$$

Under H_0, the limiting distribution of T_{mn} as $m, n \to \infty$ with $m/(m+n) \to \lambda \in (0,1)$, is normal with mean 0 and variance Kurt$(F_0) - 1$. In the introduction to Boos et al. (1989), four bootstrap resampling plans are discussed:

- I. Draw both bootstrap samples independently and with replacement from the pooled set $\{X_1, ..., X_m, Y_1, ..., Y_n\}$.
- II. Draw $X_1^*, ..., X_m^*$ with replacement from $\{X_1, ..., X_m\}$ and independently draw $Y_1^*, ..., Y_n^*$ with replacement from $\{Y_1, ..., Y_n\}$.
- III. Draw both bootstrap samples independently and with replacement from the pooled set of residuals $\{X_1 - \overline{X}, ..., X_m - \overline{X}, Y_1 - \overline{Y}, ..., Y_n - \overline{Y}\}$.
- IV. Draw $X_1^*, ..., X_m^*$ with replacement from $\{X_1/s_X, ..., X_m/s_X\}$ and independently draw $Y_1^*, ..., Y_n^*$ with replacement from $\{Y_1/s_Y, ..., Y_m/s_Y\}$.

- Plan I is not appropriate unless the means are equal, $\mu_X = \mu_Y$, an assumption that is usually not warranted. The limiting distribution of T^*_{mn} is normal with mean 0 and variance $\text{Kurt}(G) - 1$, where $G(x) = \lambda F_1(x) + (1-\lambda)F_2(x)$, and this $\text{Kurt}(G) - 1$ can be quite different from $\text{Kurt}(F_0) - 1$ when $\mu_X \neq \mu_Y$.
- For Plan II, the limiting distribution of T^*_{mn} is exactly the same as that of T_{mn} under both H_0 and any alternative. Thus, the test has approximately the correct level α under H_0, but the power is also approximately α under any alternative.

- For Plan III, the limiting distribution of T^*_{mn} is normal with mean 0 and variance $\text{Kurt}(H) - 1$, where

$$\begin{aligned} H(x) = &\quad \lambda F_1(x + \mu_X) + (1 - \lambda) F_2(x + \mu_Y) \\ = &\quad \lambda F_0(x/\sigma_X) + (1 - \lambda) F_0(x/\sigma_Y). \end{aligned}$$

  Under $H_0 : \sigma_X = \sigma_Y = \sigma$, $H(x) = F_0(x/\sigma)$ and $\text{Kurt}(H) = \text{Kurt}(F_0)$ as needed for the correct asymptotic level. Under an alternative, the kurtosis of T^*_{mn} is not the same as $\text{Kurt}(F_0)$, but because the asymptotic mean of T^*_{mn} is 0, the resulting test under an alternative haspower converging to 1.
- For Plan IV, the asymptotic results are similar to those for Plan III except that equal kurtoses for the two samples are not needed. Thus, Plan IV is more robust asymptotically, but in small samples it does not perform as well as Plan III because the pooling in Plan III produces faster convergence of critical values.

Bootstrap P-Values

Suppose that T_0 is the value of a test statistic T computed for a particular sample.

- $P(T \geq T_0 | H_0)$ is the definition of the p-value in situations where large values of T support the alternative hypothesis.
- When B resamples are made in the bootstrap world under an induced null hypothesis, define the bootstrap p-value

$$p_B = \frac{\{\#\text{ of } T_i^* \geq T_0\}}{B},$$

where $T_1^*, ..., T_B^*$ are the values of T computed from the resamples.

This is the definition we prefer and the one given by Efron and Tibshirani (1993, p. 221). However, Davison and Hinkley (1997, p. 148, 161) and others prefer $(Bp_B + 1)/(B + 1)$.

“99 Rule”

Consider a situation where the statistic T is continuous, and a parametric bootstrap gives the exact sampling distribution as B grows large.

- $T_0, T_1^*, ..., T_B^*$ are iid, all $(B+1)!$ orderings are equally likely, and p_B has a discrete uniform distribution,

$$P(p_B = 0) = P(p_B = 1/B) = ... = P(p_B = 1) = \frac{1}{1+B}.$$

- The test defined by the rejection region $p_B \leq \alpha$ has exact level α if $(B+1)\alpha$ is an integer.
- So, for small B one should use values like $B = 19$ or 39 or 99 to get standard α levels. We call this the “99 rule”.
- The “99 rule” should be followed generally in bootstrap testing situations.

Regression Settings

We consider typical regression settings based on iid random pairs $(Y_1, \boldsymbol{X}_1), \dots, (Y_n, \boldsymbol{X}_n)$, or $(Y_1, \boldsymbol{x}_1), \dots, (Y_n, \boldsymbol{x}_n)$, where the explanatory vectors $\boldsymbol{x}_i$ are viewed as fixed constants.

- In the random pairs case, it is natural to draw with replacement from the set of pairs resulting in a bootstrap resample $(Y_1^*, \boldsymbol{X}_1^*), \dots, (Y_n^*, \boldsymbol{X}_n^*)$.
- This bootstrap method is very general and applies to almost any regression method.
- The assumed model used to derive estimators does not need to be true in order for bootstrap estimates to be consistent.
- We call this method the *random pairs* bootstrap although it is really just the standard nonparametric bootstrap method.

There are a few reasons, however, to consider other bootstrap approaches in regression settings:

- Inference in regression setting is usually carried out conditional on the explanatory vectors regardless of whether they are considered fixed or random. The random pairs bootstrap, however, gives unconditional estimates.
- The random pairs bootstrap does not take advantage of any model assumptions such as an additive error structure with homogeneous errors. This nonparametric aspect of the random pairs bootstrap gives it strong robustness to model assumptions, but also can result in much less efficient procedures.

For these reasons, let us consider the *residual-based* bootstrap that is appropriate for additive errors models of the form $Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i$, where g is a known function and $e_1, ..., e_n$ are iid random errors.

- Defining the residuals as $\widehat{e}_i = Y_i - g(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})$, draw bootstrap errors $e_1^*, ..., e_n^*$ with replacement from the set

$$\left\{ (\widehat{e}_i - \overline{\widehat{e}})/\sqrt{1 - p/n}, i = 1, ..., n \right\}.$$

- Then form the bootstrap responses $Y_i^* = g(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}) + e_i^*, i = 1, ..., n$.

- If the model is linear, then the least squares estimator in the bootstrap world is $\widehat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ with variance $\text{Var}^*(\widehat{\boldsymbol{\beta}}^*) = \widehat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, where

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(\widehat{e}_i - \overline{\widehat{e}})/\sqrt{1-p/n}\right]^2 = \frac{1}{n-p}\sum_{i=1}^{n}(\widehat{e}_i - \overline{\widehat{e}})^2.$$

- Further, if the first column of $\boldsymbol{X}$ is a column of ones, then $\overline{\widehat{e}} = 0$, and we recognize the bootstrap estimate of $\text{Var}(\widehat{\boldsymbol{\beta}})$ is the same as the usual unbiased one.

Permutation and Rank Tests

- An example of two treatment effect test based on permutation method.
- six students are divided into two groups of size 2 and 4,taught with different method, the scores of students using standard method are

$$x_1 = 6, x_2 = 8$$

the scores of new method group are

$$x_3 = 7, x_4 = 18, x_5 = 11, x_6 = 9$$

As we all know,in parametric situation,we can use t-test.

In nonparametric situation, we can also use the statistic below:

$$t(X,Y) = \frac{\overline{Y} - \overline{X}}{\sqrt{s_p^2(\frac{1}{m} + \frac{1}{n})}}$$

where

$$s_p^2 = \{\sum (X_i - \overline{X})^2 + \sum (Y_i - \overline{Y})^2\}/(m+n-2)$$

If t is large,then one might be convinced that the new method is better than the standard one.

- In nonparametric situation,another common used statistic is
$$W = \text{the sum of the ranks of the } Y \text{ values}$$
when both X and Y samples are thrown together and ranked from smallest to largest.
- Let Z denote the joint sample of both X and Y together: Z=(X,Y)=(6,8,7,18,11,9),
- The ranks of these observed values are (1,3,2,6,5,4)，W=2+6+5+4=17.
- the new method is better,the W is expected to be large.
- t and W are statistics for our testing problem, we should find the distribution of each.
- There are $\binom{6}{2} = 15$ different ways of treating.

The possible samples and the values of t and W are listed below.

Table12.1 All Possible Permutations for Example Data

	X Sample		Y Sample				$\sum Y_i$	t	W
1.	6	8	7	18	11	9	45	1.17	17
2.	7	8	6	18	11	9	44	0.91	16
3.	18	8	7	6	11	9	33	-1.36	12
4.	11	8	7	18	6	9	40	0.12	13
5.	9	8	7	18	11	6	42	0.49	14
6.	6	7	8	18	11	9	46	1.47	18
7.	6	18	7	8	11	9	35	-0.84	14
8.	6	11	7	18	8	9	42	0.49	15
9.	6	9	7	18	11	9	34	-1.08	13
10.	7	18	6	8	11	9	34	-1.08	13
11.	18	11	7	6	8	9	30	-2.98	10
12.	11	9	7	18	6	8	39	-0.06	12
13.	7	11	6	18	8	9	41	0.30	14
14.	7	9	6	18	11	8	43	0.69	15
15.	18	9	7	6	11	8	32	-1.72	11

Table12.2 Permutation distribution of t

t	-2.98	-1.72	-1.36	-1.08	-0.84	-0.06	0.12
$P(t)$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$
t	0.30	0.49	0.69	0.91	1.17	1.47	
$P(t)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	

- If the treatments produce identical results,then the outcomes for each student would have been exactly the same for any of the 15 possible randomizations.
- Thus, a suitable reference distribution for t or W is just the possible 15 values of t or W along with the probability $\frac{1}{15}$ of each.
- This distribution for t or W is called the permutation distribution.

Note that the permutation distribution is discrete even when sampling from a continuous distribution.

Test Result

- A test for this experiment $(W = 17)$ with $\alpha = \frac{1}{15}$ would be reject if t$\geq$1.47, A one-sided p-value for the observed value of t=1.17 is $\frac{2}{15}$.
- Similarly, a conditional $\alpha = \frac{1}{15}$ level test based on the rank sum W would reject if W$\geq$18,and the one-sided p-value for the observed data is $\frac{2}{15}$.

- A unique feature of rank statistics when there are no ties in the data is that the permutation distribution is the same for every such data set.That is,although the data values would change in the 6 data points,the ranks would always be(1,2,3,4,5,6).
- Thus,the results for W in Table 12.1 would be exactly the same except in a different order,and therefore the distribution would be the same.
- So without ties,the exact distribution does not change.
- A standard way to rank data with ties is to assign the average rank to each of a set of tied values.
- In this case,the permutation distribution will change with samples,but the use of midranks has no effect on the general permutation method.
- But in this case, the permutation distribution can not be tabulated.

Summary:Permutation test,Rank test and Bootstrap

Permutation test

- Consider the treatment effect problem, the total N members are divided into two groups of size n and m, Computation of the permutation distribution of a test statistic involves enumeration of all $\binom{N}{n}$ divisions of the observations.This poses computational challenges.
- p-value is the proportion that are as extreme or more extreme than the observed value, we should use permutation test to obtain the exact p-value, we should calculate the test statistics $\binom{N}{n}$times.

Permutation test,Rank test and Bootstrap

- Using a part of test statistics from the permutation distribution can get an estimate of p value(Bootstrap method) and sampling can be done with replacement. If M test statistics t_i,$i = 1, ..., m$ are randomly sampled from the permutation distribution,the one-sided estimated p-value of Bootstrap is

$$\hat{p} = \frac{1 + \Sigma_{i=1}^{M} I(t_i \geq t)}{M + 1}$$

- Including the observed value t,there are M+1 test statistic values. Whereas the one-sided exact p-value of permutation test is

$$p = \frac{\Sigma_{i=1}^{\binom{N}{n}} I(t_i \geq t)}{\binom{N}{n}}$$

Permutation test,Rank test and Bootstrap

- p-value of Bootstrap is approximate
- p-value of permutation test is exact
- The rank test is a special permutation test,its test statistics does not depend on the observed data,but depend on the rank of the observed data,so it is more robust.

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

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

- 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.
- We want to control the overall false positives

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

- To control the overall false positives, we consider

$$P(making\ at\ least\ one\ false\ rejection)$$

- 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.
- Define the familywise error rate at level α if

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

- **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

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
$$\frac{\phi(x)}{x + 1/x} \le 1 - \Phi(x) \le \frac{\phi(x)}{x}, \ \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$
then approximately, the corrected bound becomes $\sigma\sqrt{2\log(2N/\alpha)/n}$. It grows like $\sqrt{\log N}$.

R code

```
rm(list=ls())
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 co
  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 #0.
mean(fwr2) #empirical familywise error for corrected test #0.06
```

False Discovery Control

- Recall we have the table:

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

- Define the *false discovery proportion* as

$$FDP = \frac{|R \cup I|}{|R|} = \frac{\#FP}{\#FP + \#TP}$$

- The *false discovery rate* is defined as the expectation of FDP. Our goal is to let

$$FDR = E[FDP] \leq \alpha$$

False Discovery Control

Benjamin-Hochberg method:

(1) Find the ordered p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$

(2) Let $j = \max\{i : p_{(i)} < i\alpha/N\}$. Let $T = p_{(j)}$.

(3) Let $R = \{i : p_i \leq T\}$.

Proof. For simplicity, assume we reject hypothesis j when $p_j \leq t$. Let $\hat{G}$ be the empirical distribution of the p-values and let $G(t) = E[\hat{G}(t)]$. In this case,

$$FDP = \frac{\sum_{i=1}^{N} I(p_i < t, H_0)}{\sum_{i=1}^{N} I(p_i < t)} = \frac{\frac{1}{N}\sum_{i=1}^{N} I(p_i < t, H_0)}{\frac{1}{N}\sum_{i=1}^{N} I(p_i < t)}.$$

Hence,

$$\begin{aligned} E[FDP] &\approx \frac{E[\frac{1}{N}\sum_{i=1}^{N} I(p_i < t, H_0)]}{\frac{1}{N}E[\sum_{i=1}^{N} I(p_i < t)]} = \frac{\frac{1}{N}\sum_{i=1}^{N} E[I(p_i < t, H_0)]}{\frac{1}{N}\sum_{i=1}^{N} E[I(p_i < t)]} \\ &= \frac{t\#P/N}{G(t)} \leq \frac{t}{G(t)} \approx \frac{t}{\hat{G}(t)} \end{aligned}$$

Let $t = p_{(i)}$ for some i; then$\hat{G}(t) = i/N$. Thus, $FDR \leq p_{(i)}N/i$. Setting it equal to α and we have the result.

Higher Criticism

- Let p_j be the p-value for test problem H_j. Then under null hypothesis, $p_j \sim Unif(0,1)$.
- Consider a level α test for an individual hypothesis test, we reject the hypothesis when $p_j < \alpha$
- Let $Y_j = I(p_j \leq \alpha)$. Under null hypothesis, $Y_j \sim Bernoulli(\alpha)$. So $\bar{Y}_n$ has mean α and standard deviation $\sqrt{\alpha(1-\alpha)/N}$
- According to CLT, let $\bar{Y}_N$ be the fraction of the rejected hypothesis with a level α test, then

$$T_N = \sqrt{N}\frac{\bar{Y}_n - \alpha}{\sqrt{\alpha(1-\alpha)}} \xrightarrow{d} N(0,1),$$

When $|T_n| > z_{\alpha/2}$, at least one hypothesis testing problem rejects H_0

Higher Criticism

- Why do we consider a fixed α? When α changes, will we get different results?
- Define the function of α. The expression is as following:
  (1) Sort p-values so that $p_{(1)} < p_{(2)} < \cdots < p_{(N)}$
  (2) Define
  $$HC_k = \sqrt{N}\frac{k/N - p_{(k)}}{\sqrt{p_{(k)}(1-p_{(k)})}}.$$
  (3) The test statistic is $T_N = \max_{1\le k\le N} HC_k$
- Note. Now $p_{(k)}$ plays the role of α. If we take $\alpha = p_{(k)}$, then the number of rejected hypothesis is k, so the fraction is k/N.
- The limiting distribution for T_n is Gumbel distribution (no need to know)

Comparison

- The three methods cares about different things
- The Bonferoni correction is to control the familywise error, which is the exact number of Type I wrong decisions. It does not need independence assumption between tests, but it may lose some power.
- The FDR method cares about the fraction of false positive. It is useful when the true positives are rare.
- Higher Criticism cares about whether there is any positive or not. It works for the case that true positives are rare, and it allows the signals to be moderately weak. On the other hand, to make sure it works, the dependence between tests cannot be strong.