



Hypothesis Testing

Professor Yongmiao Hong
Cornell University
July 7, 2020

CONTENTS

9.1 Introduction to Hypothesis Testing

9.2 Neyman-Pearson Lemma

9.3 Wald Test

9.4 Lagrangian Multiplier (LM) Test

9.5 Likelihood Ratio Test

9.6 Illustrative Examples

9.7 Conclusion

Introduction to Hypothesis Testing



Question: Given a realization \mathbf{x}^n of a random sample \mathbf{X}^n from the population distribution $f(x, \theta)$, we would also like to know whether the true parameter value θ_0 belongs to some specific subset Θ_0 of the parameter space Θ .

Introduction to Hypothesis Testing

Example 1 (9.1) [Return to Education]

Let θ measure the change in hourly wage given another year of education, holding all other factors fixed. Labor economists are interested in testing whether the return to education, controlling other factors, is zero. That is, whether or not θ equal to zero.

Introduction to Hypothesis Testing

Definition 1 (9.1) [Hypothesis]

A hypothesis is a statement about the population or some attributes of the population distribution. The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis, denoted by \mathbb{H}_0 and \mathbb{H}_A respectively.

Introduction to Hypothesis Testing

Remarks:

- A **null hypothesis** is a statement about the population or some attributes of the population. The **alternative hypothesis** is that the statement in the null hypothesis is false.
- The goal of hypothesis testing is to decide, based on an observed data set \mathbf{x}^n generated from a population, which of two complementary hypotheses is true.
- Suppose a random sample \mathbf{X}^n is generated from a population distribution $f(x, \theta)$ with some unknown value of parameter $\theta \in \Theta$, where Θ is a known finite-dimensional parameter space. In hypothesis testing, the parameter space Θ is divided into two mutually exclusive and collectively exhaustive subsets Θ_0 and Θ_A , namely $\Theta_0 \cap \Theta_A = \emptyset$ and $\Theta_0 \cup \Theta_A = \Theta$.

Introduction to Hypothesis Testing

- The problem is to determine to which of these two subsets the true value of θ belongs. That is, based upon an observed data set \mathbf{x}^n , one is trying to choose between the two hypotheses

$$\mathbb{H}_0 : \theta \in \Theta_0$$

versus

$$\mathbb{H}_A : \theta \in \Theta_A.$$

- The first hypothesis \mathbb{H}_0 is called the null hypothesis, and the second, \mathbb{H}_A , is called the alternative hypothesis.
- \mathbb{H}_0 is called the “**null**” hypothesis because it is often stated as “**no effects**” or “no relationship”. One example, is $\mathbb{H}_0 : \theta = 0$ versus $\mathbb{H}_A : \theta \neq 0$, as is the case of Example 9.1.

Introduction to Hypothesis Testing

Example 2 (9.2) [Constant Return to Scale Hypothesis]

A production function

$$Y = F(L, K)$$

tells how much output Y to produce using inputs of labor L and capital K . A production technology is said to display a constant return to scale if the output increases by the same proportion as the inputs increase; that is, for all $\lambda > 0$,

$$\lambda F(L, K) = F(\lambda L, \lambda K).$$

Introduction to Hypothesis Testing

Example 3 (9.3) [Constant Return to Scale Hypothesis]

Suppose a production function is given by

$$Y = AK^\alpha L^\beta,$$

where Y is the output, K and L are the capital and labor inputs, A is a constant, and $\theta = (\alpha, \beta)$ is a parameter vector. Then the constant return to scale hypothesis can be stated as

$$\mathbb{H}_0 : \alpha + \beta = 1.$$

The alternative hypothesis $\mathbb{H}_0 : \alpha + \beta \neq 1$ consists of two cases: $\alpha + \beta > 1$ and $\alpha + \beta < 1$, which imply an increasing return to scale and a decreasing return to scale respectively.

Introduction to Hypothesis Testing

Remarks:

- The hypotheses can be divided into two basic categories—**simple hypotheses** and **composite hypotheses**.

Introduction to Hypothesis Testing

Definition 2 (9.2) [Simple Hypothesis Versus Composite Hypothesis]

A hypothesis is *simple* if and only if it contains exactly one population. If the hypothesis contains more than one population, it is called a *composite hypothesis*.

Remarks

In Example 9.1, \mathbb{H}_0 contains only one parameter value, so \mathbb{H}_0 is a simple hypothesis. In contrast, the null hypotheses in Examples 9.2 and 9.3 contain more than one parameter values, so they are composite hypotheses.

Introduction to Hypothesis Testing

Definition 3 (9.3) [Hypothesis Testing]

A hypothesis testing procedure or a hypothesis test is a decision rule that specifies

- (1) for what sample values \mathbf{x}^n the decision is made to accept \mathbb{H}_0 as true, and
- (2) for what sample values \mathbf{x}^n \mathbb{H}_0 is rejected and \mathbb{H}_A is accepted as true.

Introduction to Hypothesis Testing

Remarks:

- The key to the construction of the decision rule for a hypothesis testing procedure is to determine the rules of rejection or acceptance of the null hypothesis.

Introduction to Hypothesis Testing

Definition 4 (9.4) [Critical Region or Rejection Region]

The set \mathbb{C} of the sample points of the random sample \mathbf{X}^n for which \mathbb{H}_0 will be rejected is called the **rejection region or critical region**. The complement of the rejection region is called the **acceptance region**.

Introduction to Hypothesis Testing

Remarks:

- A standard approach to hypothesis testing is to choose a statistic $T(\mathbf{X}^n)$ and use it to divide the sample space of \mathbf{X}^n into two mutually exclusive and exhaustive regions

$$\mathbb{A}_n(c) = \{\mathbf{x}^n : T(\mathbf{x}^n) \leq c\}$$

and

$$\mathbb{C}_n(c) = \{\mathbf{x}^n : T(\mathbf{x}^n) > c\}$$

for some prespecified constant c .

- The first region, $\mathbb{A}_n(c)$, is the acceptance region, and the second, $\mathbb{C}_n(c)$, is the rejection region. The cutoff point c is called the **critical value** and $T(\mathbf{X}^n)$ is called a test statistic.
- An important issue in hypothesis testing is to determine a suitable value of c given a data set \mathbf{x}^n . In general, we need to know the sampling distribution of $T(\mathbf{X}^n)$ under \mathbb{H}_0 in order to determine the threshold value c .

Introduction to Hypothesis Testing

Example 4 (9.4)

Suppose \mathbf{X}^n is an IID $N(\mu, \sigma^2)$ random sample, where μ is unknown but σ^2 is known. We are interested in testing for $\mathbb{H}_0 : \mu = \mu_0$ versus $\mathbb{H}_A : \mu \neq \mu_0$, where μ_0 is a known number. Here, $\Theta_0 = \{\mu_0\}$ contains one parameter value μ_0 and Θ_A contains all parameter values on the real line \mathbb{R} except μ_0 .

Introduction to Hypothesis Testing

Remarks:

- To test the null hypothesis $\mathbb{H}_0 : \mu = \mu_0$, we consider the following test statistic

$$T(\mathbf{X}^n) = \frac{\bar{X}_n - \mu_0}{\sigma / \sqrt{n}},$$

where σ is known.

- Under the null hypothesis $\mathbb{H}_0 : \mu = \mu_0$, we have

$$\begin{aligned} T(\mathbf{X}^n) &= \frac{\bar{X}_n - \mu_0}{\sigma / \sqrt{n}} \\ &\sim N(0, 1). \end{aligned}$$

Introduction to Hypothesis Testing

- Under $\mathbb{H}_A : \mu \neq \mu_0$,

$$T(\mathbf{X}^n) \sim N \left(\frac{\sqrt{n}(\mu - \mu_0)}{\sigma}, 1 \right),$$

which diverges to infinity with probability approaching one as $n \rightarrow \infty$.

- One can accept \mathbb{H}_A if $|T(\mathbf{X}^n)|$ is large. How large $T(\mathbf{X}^n)$ should be in order to be considered as “large” is determined by the sampling distribution ($N(0, 1)$) of $T(\mathbf{X}^n)$ under \mathbb{H}_0 .

Introduction to Hypothesis Testing

- Specifically, by setting $c = z_{\alpha/2}$, where $z_{\alpha/2}$ is the upper-tailed critical value of $N(0, 1)$ at level $\frac{\alpha}{2} \in (0, 1)$, i.e., $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$ where $Z \sim N(0, 1)$, we can define the acceptance and rejection regions as follows:

$$\mathbb{A}_n(c) = \left\{ \mathbf{x}^n : \left| \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \right| \leq z_{\frac{\alpha}{2}} \right\},$$
$$\mathbb{C}_n(c) = \left\{ \mathbf{x}^n : \left| \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\frac{\alpha}{2}} \right\}.$$

Introduction to Hypothesis Testing

- The $N(0, 1)$ test decision rule:
 - (1) Accept $\mathbb{H}_0 : \mu = \mu_0$ at the significance level α if
$$\mathbf{x}^n \in \mathbb{A}_n(c);$$
 - (2) Reject $\mathbb{H}_0 : \mu = \mu_0$ at the significance level α if
$$\mathbf{x}^n \in \mathbb{C}_n(c).$$
- The **significance level** α is the probability that the above decision rule will wrongly reject a correct null hypothesis. This error is called the **Type I error** and is unavoidable given any finite sample size n .

Introduction to Hypothesis Testing

Definition 5 (9.5) [Power of Test]

If \mathbb{C} is the rejection region of a test of the null hypothesis $\mathbb{H}_0 : \theta \in \Theta_0$, then the function $\pi(\theta) = P_\theta(\mathbf{X}^n \in \mathbb{C})$ is called the power of the test with the rejection region \mathbb{C} , where $P_\theta(\cdot)$ is the probability measure when the random sample follows the distribution $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$.

Introduction to Hypothesis Testing

Remarks:

- The power function $\pi(\theta)$ is the probability of rejecting \mathbb{H}_0 .
- In Example 9.4, the power of the test statistic $T(\mathbf{X}^n) = \sqrt{n}(\bar{X}_n - \mu_0)/\sigma$ is given by

$$\pi(\mu) = P \left(\left| \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2} \right).$$

Introduction to Hypothesis Testing

Definition 6 (9.6) [Type I and Type II Errors]

If $\mathbb{H}_0 : \theta \in \Theta_0$ holds and the observed data \mathbf{x}^n falls into the critical region \mathbb{C} , then a Type I error is made. The probability of making Type I error is

$$\alpha(\theta) \equiv P_{\theta}(\mathbf{X}^n \in \mathbb{C} | \mathbb{H}_0).$$

If $\mathbb{H}_A : \theta \in \Theta_0^c$ holds and the observed data \mathbf{x}^n is in the acceptance region, then a Type II error is made. The probability of making Type II error is

$$\begin{aligned} \beta(\theta) &\equiv P_{\theta}(\mathbf{X}^n \in \mathbb{A} | \mathbb{H}_A) \\ &= 1 - P_{\theta}(\mathbf{X}^n \in \mathbb{C} | \mathbb{H}_A). \end{aligned}$$

Introduction to Hypothesis Testing

Remarks:

- Under $\mathbb{H}_0 : \theta \in \Theta_0$, the power function $\pi(\theta)$ is the probability of making Type I error, namely incorrectly rejecting a correct null hypothesis.
- Type I errors are unavoidable because under \mathbb{H}_0 , because a test statistic $T(\mathbf{X}^n)$ may still take large values with nontrivial (though small) probabilities.

Introduction to Hypothesis Testing

- Under $\mathbb{H}_A : \theta \in \Theta_0^c$, the power function $\pi(\theta)$ is the probability of rejecting a false null hypothesis, and it is equal to $1 - \beta(\theta)$, where $\beta(\theta)$ is the probability of making Type II error, namely accepting an false null hypothesis.
- A test is called unbiased if $P[T(\mathbf{X}^n) > c | \mathbb{H}_A] > P[T(\mathbf{X}^n) > c | \mathbb{H}_0]$. That is, the probability to reject \mathbb{H}_0 when \mathbb{H}_0 is false is strictly larger than when it is true.

Introduction to Hypothesis Testing

- There exists a tradeoff between Type I errors and Type II errors given any sample size n . For any given n , if the critical region \mathbb{C} shrinks, the probability of making Type I error decreases, but the probability of making Type II error increases. Similarly, if the critical region \mathbb{C} increases, the Type II error $\beta(\theta)$ decreases, but Type I error $\alpha(\theta)$ increases.
- Usually, hypothesis tests are evaluated and compared through their probabilities of making mistakes. The classical approach to hypothesis testing is to bound the probability of Type I error by some value $\alpha \in (0, 1)$ over all values of θ in Θ_0 and to try to find a test that minimizes the probability of a Type II error over all values of θ in Θ_A .

Introduction to Hypothesis Testing

Definition 7 (9.7) [Uniformly Most Powerful Test]

Let \mathbb{T} be a class of tests for testing $\mathbb{H}_0 : \theta \in \Theta_0$ versus $\mathbb{H}_A : \theta \in \Theta_A$. A test $T(\mathbf{X}^n)$ in class \mathbb{T} , with power function $\pi(\theta)$, is a uniformly most powerful test over \mathbb{T} if $\pi(\theta) \geq \tilde{\pi}(\theta)$ for all $\theta \in \Theta_A$, where $\tilde{\pi}(\theta)$ is the power function of any other test $G(\mathbf{X}^n)$ in class \mathbb{T} .

Introduction to Hypothesis Testing

Remarks:

- Suppose for a test statistic $T(\mathbf{X}^n)$, $P[T(\mathbf{X}^n) > c | \mathbb{H}_0] \leq \alpha$. Then the value of α gives the maximum Type I error for the test statistic $T(\mathbf{X}^n)$, and is called the level of the test.
- If $T(\mathbf{X}^n)$ has level α and $P[T(\mathbf{X}^n) > c | \mathbb{H}_0] = \alpha$, then the test is called a size α test. Obviously, the class of level α tests contains the set of size α tests.

Introduction to Hypothesis Testing

- Usually, we will choose \mathbb{T} to be a class of tests with the same level or same size α . In some complicated testing situations, however, it may be computationally impossible to construct a size α test. In such situations, a researcher must be satisfied with a level α test, realizing that some compromises may be made.
- An example is the so-called Bonferroni's method, which is used to address the problem of multiple comparisons when we have k individual tests for the null hypothesis \mathbb{H}_0 :

Introduction to Hypothesis Testing

- Suppose for each $i \in \{1, \dots, k\}$, $T_i(\mathbf{X}^n)$ is a level α/k test for \mathbb{H}_0 , namely

$$P [T_i(\mathbf{X}^n) > c_i | \mathbb{H}_0] \leq \frac{\alpha}{k}, \quad i = 1, \dots, k,$$

where c_i is a critical value for $T_i(\mathbf{X}^n)$.

Introduction to Hypothesis Testing

- Then the maximum test statistic

$$T(\mathbf{X}^n) = \max_{1 \leq i \leq k} T_i(\mathbf{X}^n)$$

is a level α test, because

$$\begin{aligned} P[T(\mathbf{X}^n) > c | \mathbb{H}_0] &= P \left[\max_{1 \leq i \leq k} T_i(\mathbf{X}^n) > c | \mathbb{H}_0 \right] \\ &= P \left[\bigcup_{i=1}^k \{T_i(\mathbf{X}^n) > c\} | \mathbb{H}_0 \right] \\ &\leq \sum_{i=1}^k P [T_i(\mathbf{X}^n) > c | \mathbb{H}_0] = \alpha, \end{aligned}$$

where the inequality follows by Boole's inequality.

Introduction to Hypothesis Testing

Remarks:

- To test a economic hypothesis, we have to transform it into a statistical hypothesis, and then test the statistical hypothesis using an observed economic data.
- In transforming an economic hypothesis into a statistical hypothesis, some auxiliary assumptions are often imposed. This induces a gap between the original economic hypothesis and the resulting statistical hypothesis.
- This gap may cause some problem in economic interpretation of the empirical results on testing the statistical hypothesis

Introduction to Hypothesis Testing

Example 5 (9.5)[Efficient Market Hypothesis]

Suppose R_t is the return on some asset or portfolio in time period t , and $I_{t-1} = (R_{t-1}, R_{t-2}, \dots)$ denotes the historical asset return information available at time $t - 1$. The asset market is called informationally weakly efficient if

$$E(R_t | I_{t-1}) = E(R_t).$$

That is, the historical asset return information has no predictive power for future asset return.

To test this economic hypothesis, one can consider a linear autoregressive model

Introduction to Hypothesis Testing

Example 5 (9.5)[Efficient Market Hypothesis]

$$R_t = \alpha_0 + \sum_{j=1}^k \alpha_j R_{t-j} + \varepsilon_t,$$

where ε_t is a stochastic disturbance. Under the efficient market hypothesis (EMH), we have

$$\mathbb{H}_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0.$$

Introduction to Hypothesis Testing

Example 5 (9.5)[Efficient Market Hypothesis]

If one has evidence that at least one α_j , $j \in \{1, \dots, k\}$, is not zero, it will imply that the efficient market hypothesis holds.

However, when one does not reject the statistical hypothesis \mathbb{H}_0 , this does not necessarily imply that the original economic hypothesis—the efficient market hypothesis holds. The reason is that the linear autoregressive model is just one of many (possibly infinite) ways to test predictability of the historical asset returns for future asset return. In other words, the predictability may arise in a nonlinear manner. Therefore, there exists a gap between the efficient market hypothesis and the statistical hypothesis \mathbb{H}_0 . Because of this gap, when one does not reject \mathbb{H}_0 , one can only say that no evidence against the efficient market hypothesis is found rather than conclude that the efficient market hypothesis holds.

Introduction to Hypothesis Testing

Remarks:

- There exists a gap between **data predictability** versus **model predictability**.

CONTENTS

9.1 Introduction to Hypothesis Testing

9.2 Neyman-Pearson Lemma

9.3 Wald Test

9.4 Lagrangian Multiplier (LM) Test

9.5 Likelihood Ratio Test

9.6 Illustrative Examples

9.7 Conclusion

Neyman-Pearson Lemma

Theorem 1 (9.1) [Neyman-Pearson Lemma]

Consider testing a simple null hypothesis $\mathbb{H}_0 : \theta = \theta_0$ versus a simple alternative hypothesis $\mathbb{H}_A : \theta = \theta_1$, where the PMF/PDF of the random sample \mathbf{X}^n corresponding to $\theta_i, i \in \{0, 1\}$, is $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_i)$. Suppose a test with rejection and acceptance regions $\mathbb{C}_n(c)$ and $\mathbb{A}_n(c)$ respectively is defined as follows:

Neyman-Pearson Lemma

Theorem 1 (9.1) [Neyman-Pearson Lemma]

(a)

$$\mathbb{C}_n(c) = \left\{ \mathbf{x}^n : \frac{f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1)}{f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)} > c \right\}$$

and

$$\mathbb{A}_n(c) = \left\{ \mathbf{x}^n : \frac{f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1)}{f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)} \leq c \right\}$$

for some constant $c \geq 0$, and

Neyman-Pearson Lemma

Theorem 1 (9.1) [Neyman-Pearson Lemma]

(b)

$$P[\mathbf{X}^n \in \mathbb{C}_n(c) | \mathbb{H}_0] = \alpha.$$

Then

(1) [**Sufficiency**] Any test that satisfies conditions (a) and (b) is a uniformly most powerful level α test.

(2) [**Necessity**] If there exists a test satisfying conditions (a) and (b) with $c > 0$, then every uniformly most powerful level α test is a size α test (i.e., satisfying condition (b)), and every uniformly most powerful level α test satisfies condition (a) except perhaps on a set A in the sample space of \mathbf{X}^n satisfying $P(\mathbf{X}^n \in A | \mathbb{H}_0) = P(\mathbf{X}^n \in A | \mathbb{H}_A) = 0$.

Neyman-Pearson Lemma

Proof:

- Note that

$$\begin{aligned} P[\mathbf{X}^n \in \mathbb{C}_n(c) | \mathbb{H}_0] &= E \{ \mathbf{1}[\mathbf{X}^n \in \mathbb{C}_n(c)] | \mathbb{H}_0 \} \\ &= \int \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] f(\mathbf{x}^n, \theta_0) d\mathbf{x}^n, \end{aligned}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

To be Continued

Neyman-Pearson Lemma

Proof:

- (1) We first show that a test (denoted as $T(\mathbf{X}^n)$) that satisfies conditions (a) and (b) is uniformly most powerful.

Suppose there is another test (denoted $T_1(\mathbf{X}^n)$) with $E\{\mathbf{1}[\mathbf{X}^n \in \mathbb{C}_{1n}] | \mathbb{H}_0\} \leq \alpha$. (The test $T_1(\mathbf{X}^n)$ need not be a likelihood ratio test.) We shall show that $T_1(\mathbf{X}^n)$ is not more powerful than $T(\mathbf{X}^n)$.

Observe that if $\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] > \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{1n}]$, then the sample point \mathbf{x}^n is in the critical region $\mathbb{C}_n(c)$ of the test $T(\mathbf{X}^n)$, and so $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) > cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)$; on the other hand, if $\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] < \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{1n}]$, then the sample point \mathbf{x}^n is in the acceptance region $\mathbb{A}_n(c)$ of the test $T(\mathbf{X}^n)$, and so $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) \leq cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)$. In either case, we have:

$$\{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{1n}]\} [f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) - cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)] \geq 0$$

To be Continued

Neyman-Pearson Lemma

Proof:

Thus,

$$\int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{1n}]\} [f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) - c f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)] d\mathbf{x}^n \geq 0$$

This implies

$$\begin{aligned} & \int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{1n}]\} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) d\mathbf{x}^n \\ & \geq c \int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{1n}]\} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n. \end{aligned}$$

To be Continued

Neyman-Pearson Lemma

Proof:

Because $\int_{\mathbb{R}^n} \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{1n}] f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n \leq \alpha$, $\int_{\mathbb{R}^n} \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n = \alpha$, and $c \geq 0$, we have

$$c \int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{1n}]\} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n \geq 0.$$

It follows that

$$\int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{1n}]\} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) d\mathbf{x}^n \geq 0,$$

which implies

$$P[\mathbf{X}^n \in \mathbb{C}_n(c) | \mathbb{H}_A] \geq P[\mathbf{X}^n \in \mathbb{C}_{1n} | \mathbb{H}_A].$$

That is, the test $T_1(\mathbf{X}^n)$ is not more powerful than $T(\mathbf{X}^n)$.

To be Continued

Neyman-Pearson Lemma

Proof:

- (2) Assume that $T(\mathbf{X}^n)$ is a test that satisfies Conditions (a) and (b) with $c > 0$.
- (i) We first show that a uniformly most powerful level α test (denoted as $T_2(\mathbf{X}^n)$) is a size α test (i.e., satisfies Condition (b)).

Suppose it is not a size α test. Then $\int_{\mathbb{R}^n} \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{2n}] f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n < \alpha$. Given Condition (b) for the test $T(\mathbf{X}^n)$, we have $\int_{\mathbb{R}^n} \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n = \alpha$. It follows that

$$\int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{2n}]\} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n > 0.$$

To be Continued

Neyman-Pearson Lemma

Proof:

Observe that, if $\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{2n}] > 0$, then \mathbf{x}^n is in the rejection region $\mathbb{C}_n(c)$ of the test $T(\mathbf{X}^n)$, and so $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) > cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)$; if $\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{2n}] < 0$, then \mathbf{x}^n is in the acceptance region $\mathbb{A}_n(c)$ of the test $T(\mathbf{X}^n)$, and so $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) \leq cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)$. It follows that

$$\int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{2n}]\} [f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) - cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)] d\mathbf{x}^n \geq 0.$$

Therefore,

$$\begin{aligned} & \int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{2n}]\} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) d\mathbf{x}^n \\ & \geq c \int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_{2n}]\} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n \\ & > 0 \end{aligned}$$

given $c > 0$. This implies $P[\mathbf{X}^n \in \mathbb{C}_n(c) | \mathbb{H}_A] > P[\mathbf{X}^n \in \mathbb{C}_{2n} | \mathbb{H}_A]$, suggesting that the test $T_2(\mathbf{X}^n)$ is not uniformly most powerful, a contradiction.

To be Continued

Neyman-Pearson Lemma

Proof:

- (ii) Next, we show that every uniformly most powerful level α test must satisfy Conditions (a) and (b) except perhaps on a zero probability set A in the sample space of \mathbf{X}^n satisfying $P(\mathbf{X}^n \in A | \mathbb{H}_0) = P(\mathbf{X}^n \in A | \mathbb{H}_A) = 0$. Suppose $T^*(\mathbf{X}^n)$ is a most powerful level α test with the rejection region \mathbb{C}_n^* . Because every most powerful test with level α is a size α test, we have

$$\int_{\mathbb{R}^n} \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n^*] f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n = \alpha = \int_{\mathbb{R}^n} \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) d\mathbf{x}^n.$$

Also, both $T(\mathbf{X}^n)$ and $T^*(\mathbf{X}^n)$ are most powerful tests, so they are equally powerful under \mathbb{H}_A , namely

$$\int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n^*]\} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) d\mathbf{x}^n = 0.$$

To be Continued

Neyman-Pearson Lemma

Proof:

Thus

$$\int_{\mathbb{R}^n} \{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n^*]\} [f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) - cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)] d\mathbf{x}^n = 0.$$

Now, observe that if $\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n^*] > 0$, then \mathbf{x}^n must be in the rejection region $\mathbb{C}_n(c)$ of the test $T(\mathbf{X}^n)$, and so $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) - cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) > 0$; if $\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n^*] < 0$, then \mathbf{x}^n must be in the acceptance region of the test $T(\mathbf{X}^n)$, and so $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) - cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0) \leq 0$. In both cases, the product $\{\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n^*]\} [f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1) - cf_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)] \geq 0$. Given that the integral of this nonnegative product is 0, we have that $\mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n(c)] - \mathbf{1}[\mathbf{x}^n \in \mathbb{C}_n^*] = 0$ for all \mathbf{x}^n in the sample space of \mathbf{X}^n except for a set with zero probability. Therefore, the sets \mathbb{C}_n^* and $\mathbb{C}_n(c)$ are identical except on the set of probability 0.

Neyman-Pearson Lemma

Remarks:

- The Neyman-Pearson lemma provides a method of finding a uniformly most powerful test when both the null and the alternative are simple hypotheses.
- The **likelihood ratio** test is the uniformly most powerful test.
- However, when a hypothesis is composite, i.e., the hypothesis contains more than one parameter value, the lemma may not hold. See Hong and Lee (2013) for an example.

Neyman-Pearson Lemma

Corollary 1 (9.2) [Likelihood Ratio Test and Sufficient Statistics]

Suppose $T(\mathbf{X}^n)$ is a sufficient statistic for θ and $g(t, \theta_i)$ is the PMF/PDF of $T(\mathbf{X}^n)$ corresponding to $\theta_i, i \in \{0, 1\}$. Then any test based on $T(\mathbf{X}^n)$ with rejection region \mathbb{R}_n is a uniformly most powerful level α test for $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_A : \theta = \theta_1$ if the test has the rejection and acceptance regions

$$\mathbb{C}_n(c) = \left\{ t : \frac{g(t, \theta_1)}{g(t, \theta_0)} > c \right\}$$

and

$$\mathbb{A}_n(c) = \left\{ t : \frac{g(t, \theta_1)}{g(t, \theta_0)} \leq c \right\}$$

for some $c \geq 0$, where $P[T(\mathbf{X}^n) \in \mathbb{C}_n(c) | \mathbb{H}_0] = \alpha$.

Neyman-Pearson Lemma

Remarks:

- The likelihood ratio test based on the random sample X^n can be reduced to a likelihood ratio test based on the sufficient statistic $T(\mathbf{X}^n)$ of θ which remains to be the uniformly most powerful test.

Neyman-Pearson Lemma

Example 6 (9.6)

Suppose \mathbf{X}^n is an IID random sample from an Exponential(θ) PDF $f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}$ for $x \geq 0$. Find a uniformly most powerful level α test for $\mathbb{H}_0 : \theta = 1$ versus $\mathbb{H}_A : \theta = 2$.

Neyman-Pearson Lemma

Solution

- The likelihood function of the random sample \mathbf{X}^n when \mathbf{x}^n is observed is

$$\begin{aligned} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) &= \prod_{i=1}^n f(x_i, \theta) \\ &= \frac{1}{\theta^n} e^{-n\bar{x}_n/\theta}, \end{aligned}$$

where $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$.

- By the factorization theorem in Theorem 6.14, \bar{X}_n is a sufficient statistic for θ . Since the sum of IID exponential(θ) random variables follows a Gamma(n, θ) distribution, and so the sample mean $\bar{X}_n \sim \text{Gamma}(n, \frac{\theta}{n})$. Thus, its PDF

$$g(\bar{x}_n, \theta) = \frac{n^n}{(n-1)! \theta^n} \bar{x}_n^{n-1} e^{-n\bar{x}_n/\theta}, \text{ for } \bar{x}_n > 0.$$

To be Continued

Neyman-Pearson Lemma

- It follows that the likelihood ratio

$$\begin{aligned}\frac{f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_1)}{f_{\mathbf{X}^n}(\mathbf{x}^n, \theta_0)} &= \frac{g(\bar{x}_n, \theta_1)}{g(\bar{x}_n, \theta_0)} \\ &= \frac{\frac{1}{2^n} e^{-\frac{n}{2} \bar{x}_n}}{e^{-n \bar{x}_n}} \\ &= \frac{1}{2^n} e^{\frac{n}{2} \bar{x}_n}.\end{aligned}$$

- Define a test with the following one-dimensional rejection region

$$\bar{x}_n \in \mathbb{C}_n(c) \text{ if } \frac{1}{2^n} e^{\frac{n}{2} \bar{x}_n} > c,$$

or equivalently,

$$\bar{x}_n \in \mathbb{C}_n(c) \text{ if } \bar{x}_n > 2 \ln 2 + \frac{2 \ln c}{n}.$$

To be Continued

Neyman-Pearson Lemma

- Then we determine the value of c so that the test has size α . This requires

$$\begin{aligned}\alpha &= \int_{2 \ln 2 + 2n^{-1} \ln c}^{\infty} g(\bar{x}_n, \theta_0) d\bar{x}_n \\ &= \int_{2 \ln 2 + 2n^{-1} \ln c}^{\infty} \frac{n^n}{(n-1)!} \bar{x}_n^{n-1} e^{-n\bar{x}_n} d\bar{x}_n.\end{aligned}$$

Solving this nonlinear equation, we can obtain $c = c(\alpha, n)$ as a function of α and n but it has no closed form solution. By the Neyman-Pearson lemma and Corollary 9.2, the above test is the uniformly most powerful size α test.

To be Continued

CONTENTS

9.1 Introduction to Hypothesis Testing

9.2 Neyman-Pearson Lemma

9.3 Wald Test

9.4 Lagrangian Multiplier (LM) Test

9.5 Likelihood Ratio Test

9.6 Illustrative Examples

9.7 Conclusion

Wald Test

- Consider the hypotheses of interest

$$\mathbb{H}_0 : g(\theta) = \mathbf{0}$$

versus

$$\mathbb{H}_A : g(\theta) \neq \mathbf{0}$$

where $g : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is a continuously differentiable q -dimensional vector-valued function of a p -dimensional parameter vector θ .

- The integer q is the number of restrictions on parameter vector θ . We assume $q \leq p$.
- One important example is a linear vector-valued function

$$g(\theta) = R\theta - r,$$

where R is a $q \times p$ known constant matrix, r is a $q \times 1$ known constant vector. The null hypothesis

$$\mathbb{H}_0 : R\theta = r$$

imposes q linear restrictions on the p -dimensional parameter vector θ .

Wald Test

- **Regularity conditions**
- **Assumption W.1 [Asymptotic Normality]:** $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$, where V is a $p \times p$ symmetric bounded and nonsingular matrix, θ_0 is the true parameter value which is an interior point in Θ , and Θ is a compact parameter space.
- **Assumption W.2 [Consistent Variance Estimator]:** $\hat{V} \xrightarrow{p} V$ as $n \rightarrow \infty$.
- **Assumption W.3 [Smooth Condition on Restriction Function]:** $g : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is a continuously differentiable function of $\theta \in \Theta$, and the $q \times p$ matrix $G(\theta_0) = \frac{\partial}{\partial \theta} g(\theta_0)$ has rank q , where $q \leq p$.

Wald Test

Remarks:

- Assumption W.1 allows the estimator $\hat{\theta}$ to be any root- n consistent asymptotically normal estimator.
- Assumption W.2 assumes that there exists a consistent estimator \hat{V} for the asymptotic variance V of $\sqrt{n}(\hat{\theta} - \theta_0)$. Suppose $\hat{\theta}$ admits the following asymptotic expansion

$$\sqrt{n}(\hat{\theta} - \theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^n \psi(X_i, \theta_0) + o_P(1),$$

for some function $\psi(X_i, \theta_0)$, where $\mathbf{X}^n = (X_1, \dots, X_n)$ is an IID sequence and $E[\psi(X_i, \theta_0)] = 0$, where the expectation $E(\cdot)$ is taken under the population distribution $f_X(x) = f(x, \theta_0)$. Then we have

$$V = E[\psi(X_i, \theta_0)\psi(X_i, \theta_0)'].$$

Therefore, a consistent estimator can be given by

$$\hat{V} = \frac{1}{n} \sum_{i=1}^n \psi(X_i, \hat{\theta})\psi(X_i, \hat{\theta})'.$$

By ULLN for an IID random sample, primitive regularity conditions can be provided to ensure $\hat{V} \rightarrow V$ as $n \rightarrow \infty$ almost surely.

Wald Test

- As an example, if $\hat{\theta}$ is an MLE, then from the proof of the asymptotic normality of the MLE, we have

$$\psi_i(X_i, \theta_0) = -H^{-1}(\theta_0) \frac{\partial \ln f(X_i, \theta_0)}{\partial \theta}.$$

It follows that

$$\begin{aligned} V &= E[\psi(X_i, \theta_0)\psi(X_i, \theta_0)'] \\ &= H^{-1}(\theta_0)I(\theta_0)H^{-1}(\theta_0) \\ &= -H^{-1}(\theta_0) \end{aligned}$$

where the last equality follows from the information matrix equality $I(\theta_0) + H(\theta_0) = \mathbf{0}$, where both the information matrix $I(\theta)$ and the Hessian matrix $H(\theta)$.

- Assumption W.3 is a regularity condition on the restriction function $g(\cdot)$. The full rank condition for the $q \times p$ matrix $G(\theta_0)$ and $q \leq p$ ensure that the $q \times q$ symmetric matrix $G(\theta_0)VG(\theta_0)'$ is nonsingular.

Wald Test

- To test $\mathbb{H}_0 : g(\theta_0) = 0$, a natural approach is to base a test on statistic $g(\hat{\theta})$, where $\hat{\theta}$ is a consistent estimator of θ_0 . Because $g(\cdot)$ is continuous, we always have $g(\hat{\theta}) \xrightarrow{p} g(\theta_0)$ whenever $\hat{\theta} \xrightarrow{p} \theta_0$ as $n \rightarrow \infty$. It follows that $g(\hat{\theta})$ will be close to zero under \mathbb{H}_0 and will converge to a nonzero limit under \mathbb{H}_A . Thus, we can test \mathbb{H}_0 by checking whether $g(\hat{\theta})$ is close to zero.
- How large the value of $g(\hat{\theta})$ should be in order to be considered as significantly different from zero will be determined by the sampling distribution of $g(\hat{\theta})$ under \mathbb{H}_0 .
- By the mean value theorem, we have

$$g(\hat{\theta}) = g(\theta_0) + G(\bar{\theta}) \left(\hat{\theta} - \theta_0 \right),$$

where $\bar{\theta} = \lambda \hat{\theta} + (1 - \lambda)\theta_0$ for some $\lambda \in [0, 1]$, and the gradient function

$$G(\theta) = \frac{dg(\theta)}{d\theta}$$

is a $q \times p$ matrix.

Wald Test

- Given that $\|\bar{\theta} - \theta_0\| = \|\lambda(\hat{\theta} - \theta_0)\| \leq \|\hat{\theta} - \theta_0\| \xrightarrow{p} 0$ as $n \rightarrow \infty$ and the continuity of $G(\cdot)$, we have $G(\bar{\theta}) \xrightarrow{p} G(\theta_0)$ as $n \rightarrow \infty$. Then by the asymptotic normality that $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$, and the Slutsky theorem, we have

$$\sqrt{n} \left[g(\hat{\theta}) - g(\theta_0) \right] \xrightarrow{d} N(0, G(\theta_0)VG(\theta_0)').$$

Under $\mathbb{H}_0 : g(\theta_0) = 0$, we have

$$\sqrt{n}g(\hat{\theta}) \xrightarrow{d} N(0, G(\theta_0)VG(\theta_0)').$$

Wald Test

Since the $q \times q$ matrix $G(\theta_0)VG(\theta_0)'$ is nonsingular given the full rank condition on $G(\theta_0)$ and the nonsingularity condition on V , the quadratic form

$$\begin{aligned} & \sqrt{n}g(\hat{\theta})' [G(\theta_0)VG(\theta_0)']^{-1} \sqrt{n}g(\hat{\theta}) \\ & \xrightarrow{d} \chi_q^2. \end{aligned}$$

Because $G(\hat{\theta}) \xrightarrow{p} G(\theta_0)$ as $n \rightarrow \infty$ by continuity of $G(\cdot)$ and $\hat{\theta} \xrightarrow{p} \theta_0$, and $\hat{V} \xrightarrow{p} V$ by Assumption W.2, we have

$$G(\hat{\theta})\hat{V}G(\hat{\theta})' \xrightarrow{p} G(\theta_0)VG(\theta_0)'.$$

Therefore, the stochastic matrix $G(\hat{\theta})\hat{V}G(\hat{\theta})'$ is nonsingular for n sufficiently large, and by the Slutsky theorem, the Wald test statistic

$$\begin{aligned} W &= n \cdot g(\hat{\theta})' \left[G(\hat{\theta})\hat{V}G(\hat{\theta})' \right]^{-1} g(\hat{\theta}) \\ & \xrightarrow{d} \chi_q^2 \end{aligned}$$

under \mathbb{H}_0 .

Wald Test

Theorem 3 (9.3) [Wald Test]

Suppose Assumptions W.1–W.3 and \mathbb{H}_0 hold. Then as $n \rightarrow \infty$,

$$W = n \cdot g(\hat{\theta})' \left[G(\hat{\theta}) \hat{V} G(\hat{\theta})' \right]^{-1} g(\hat{\theta}) \xrightarrow{d} \chi_q^2.$$

Wald Test

Remarks:

- \hat{W} is a quadratic form in the difference $\sqrt{n}g(\hat{\theta})$ and $\sqrt{n}g(\theta_0) = \mathbf{0}$, weighted by the asymptotic variance estimator $G(\hat{\theta})\hat{V}G(\hat{\theta})'$ of $\sqrt{n}[g(\hat{\theta}) - g(\theta_0)]$.
- The Wald test is an asymptotically size α test that rejects $\mathbb{H}_0 : g(\theta_0) = \mathbf{0}$ when \hat{W} exceeds the $(1 - \alpha)$ th quantile of the χ_q^2 distribution.
- Under $\mathbb{H}_A : g(\theta_0) \neq \mathbf{0}$, we have $g(\hat{\theta}) \xrightarrow{p} g(\theta_0) \neq \mathbf{0}$, $G(\hat{\theta}) \xrightarrow{p} G(\theta_0)$, and $\hat{V} \xrightarrow{p} V$. It follows that

$$\frac{W}{n} \rightarrow g(\theta_0)' [G(\theta_0)VG(\theta_0)']^{-1} g(\theta_0) > 0$$

under \mathbb{H}_A . In other words, with probability approaching one, the Wald statistic W diverges to positive infinity at the rate of n , thus ensuring asymptotic power one of the test under \mathbb{H}_A at any given significance level $\alpha \in (0, 1)$.

Wald Test

- \hat{W} is applicable for many root- n consistent estimators $\hat{\theta}$. We now consider a special but important case: that is, when $\hat{\theta}$ is an MLE. In this case, $V = -H^{-1}(\theta_0)$, and we can use the asymptotic variance estimator $\hat{V} = [-\hat{H}(\hat{\theta})]^{-1}$, where the sample Hessian matrix

$$\hat{H}(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \ln f(X_i, \theta)}{\partial \theta \partial \theta'}.$$

The resulting Wald test statistic can be constructed as follows:

$$\hat{W} = ng(\hat{\theta})' \left[-G(\hat{\theta}) \hat{H}^{-1}(\hat{\theta}) G(\hat{\theta})' \right]^{-1} g(\hat{\theta}).$$

If in addition the regularity conditions of Section 8.3 hold, we can show $\hat{H}(\hat{\theta}) \rightarrow H(\theta_0)$ as $n \rightarrow \infty$ almost surely, and therefore $W \rightarrow \chi_q^2$ under \mathbb{H}_0 .

Note that the Wald test statistic W only involves estimation under the alternative \mathbb{H}_A .

CONTENTS

9.1 Introduction to Hypothesis Testing

9.2 Neyman-Pearson Lemma

9.3 Wald Test

9.4 Lagrangian Multiplier (LM) Test

9.5 Likelihood Ratio Test

9.6 Illustrative Examples

9.7 Conclusion

Lagrangian Multiplier (LM) Test

- Suppose we have an IID random sample \mathbf{X}^n from the population $f(x, \theta_0)$, where θ_0 is an unknown parameter value in Θ .
- Consider the normalized log-likelihood

$$\hat{l}(\theta) = \frac{1}{n} \sum_{i=1}^n \ln f(X_i, \theta),$$

and the constrained maximum likelihood estimator that solves the constrained maximization problem

$$\tilde{\theta} = \arg \max_{\theta \in \Theta} \hat{l}(\theta)$$

subject to the constraint that $g(\theta) = \mathbf{0}$.

- Define the Lagrangian function

$$L(\theta, \lambda) = \hat{l}(\theta) + \lambda' g(\theta),$$

where λ is the Lagrangian multiplier. Let $\tilde{\lambda}$ be the corresponding maximizing value of λ .

Lagrangian Multiplier (LM) Test

- Then the first order conditions (FOC) are

$$\begin{aligned}\frac{\partial L(\tilde{\theta}, \tilde{\lambda})}{\partial \theta} &= \frac{\partial \hat{l}(\tilde{\theta})}{\partial \theta} + G(\tilde{\theta})' \tilde{\lambda} = \mathbf{0}, \\ \frac{\partial L(\tilde{\theta}, \tilde{\lambda})}{\partial \lambda} &= g(\tilde{\theta}) = \mathbf{0},\end{aligned}$$

- By the mean value theorem, we have

$$\begin{aligned}G(\tilde{\theta})' \tilde{\lambda} &= -\frac{d\hat{l}(\tilde{\theta})}{d\theta} \\ &= -\frac{d\hat{l}(\theta_0)}{d\theta} - \frac{d^2\hat{l}(\bar{\theta}_a)}{d\theta d\theta'} (\tilde{\theta} - \theta_0),\end{aligned}$$

Lagrangian Multiplier (LM) Test

where $\bar{\theta}_a = a\tilde{\theta} + (1-a)\theta_0$ for some $a \in [0, 1]$, which lies on the segment between $\tilde{\theta}$ and θ_0 . Note

$$\frac{d^2\hat{l}(\theta)}{d\theta d\theta'} = \hat{H}(\theta)$$

is the sample Hessian matrix. Given the regularity conditions in Section 8.3, we have shown there that $\hat{H}(\hat{\theta}) \rightarrow H(\theta_0)$ as $n \rightarrow \infty$ almost surely for any consistent estimator $\hat{\theta}$ of θ_0 . Because $H(\theta_0)$ is nonsingular, $\hat{H}^{-1}(\hat{\theta}) \rightarrow H^{-1}(\theta_0)$ as $n \rightarrow \infty$ almost surely, and $\hat{H}^{-1}(\hat{\theta})$ exists for n sufficiently large. It follows that

$$\hat{H}(\bar{\theta}_a)^{-1}G(\tilde{\theta})'\tilde{\lambda} = -\hat{H}(\bar{\theta}_a)^{-1}\frac{d\hat{l}(\theta_0)}{d\theta} - (\tilde{\theta} - \theta_0), \quad (9.1)$$

Lagrangian Multiplier (LM) Test

- Next, by the mean value theorem again, we have

$$\mathbf{0} = g(\tilde{\theta}) = g(\theta_0) + G(\bar{\theta}_b)(\tilde{\theta} - \theta_0),$$

where $\bar{\theta}_b = b\tilde{\theta} + (1 - b)\theta_0$ for some $b \in [0, 1]$, which lies on the segment between $\tilde{\theta}$ and θ_0 . It follows that under $\mathbb{H}_0 : g(\theta_0) = 0$, we have

$$G(\bar{\theta}_b)(\tilde{\theta} - \theta_0) = \mathbf{0}. \quad (9.2)$$

Hence, multiplying Eq.(9.1) by $G(\bar{\theta}_b)$ and using Eq.(9.2), we obtain

$$\begin{aligned} & G(\bar{\theta}_b)\hat{H}(\bar{\theta}_a)^{-1}G(\tilde{\theta})'\tilde{\lambda} \\ = & -G(\bar{\theta}_b)\hat{H}(\bar{\theta}_a)^{-1}\frac{d\hat{l}(\theta_0)}{d\theta} - G(\bar{\theta}_b)(\tilde{\theta} - \theta_0) \\ = & -G(\bar{\theta}_b)\hat{H}(\bar{\theta}_a)^{-1}\frac{d\hat{l}(\theta_0)}{d\theta}. \end{aligned}$$

Lagrangian Multiplier (LM) Test

By the continuity of function $G(\cdot)$, $\|\bar{\theta}_a - \theta_0\| \leq \|\tilde{\theta} - \theta_0\| \rightarrow 0$, $\|\bar{\theta}_b - \theta_0\| \leq \|\tilde{\theta} - \theta_0\| \rightarrow 0$ and $\hat{H}(\hat{\theta}) \rightarrow H(\theta_0)$ as $n \rightarrow \infty$ almost surely, we have

$$G(\bar{\theta}_b)\hat{H}(\bar{\theta}_a)^{-1}G(\tilde{\theta})' \rightarrow G(\theta_0)H^{-1}(\theta_0)G(\theta_0)' \text{ a.s.}$$

and the latter is nonsingular. Therefore, for n sufficiently large, $G(\bar{\theta}_b)\hat{H}(\bar{\theta}_a)^{-1}G(\tilde{\theta})'$ is nonsingular as well. Hence,

$$\begin{aligned} \sqrt{n}\tilde{\lambda} &= - \left[G(\bar{\theta}_b)\hat{H}(\bar{\theta}_a)^{-1}G(\tilde{\theta})' \right]^{-1} G(\bar{\theta}_b)\hat{H}(\bar{\theta}_a)^{-1} \sqrt{n} \frac{d\hat{l}(\theta_0)}{d\theta} \\ &= -\hat{A} \sqrt{n} \frac{d\hat{l}(\theta_0)}{d\theta}, \text{ say.} \end{aligned}$$

Lagrangian Multiplier (LM) Test

- By CLT for an IID random sequence, we have

$$\sqrt{n} \frac{d\hat{l}(\theta_0)}{d\theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \ln f(X_i, \theta_0)}{\partial \theta} \xrightarrow{d} N(0, I(\theta_0)),$$

where $I(\theta_0)$ is Fisher's information matrix evaluated at $\theta = \theta_0$.

- On the other hand,

$$\hat{A} \xrightarrow{p} [G(\theta_0)H(\theta_0)^{-1}G(\theta_0)']^{-1} G(\theta_0)H(\theta_0) = A_0, \text{ say.}$$

It follows from the Slutsky theorem that

$$\begin{aligned} & \sqrt{n}\tilde{\lambda} \xrightarrow{d} N(0, A_0 I(\theta_0) A_0') \\ & \sim N(0, -[G(\theta_0)H(\theta_0)^{-1}G(\theta_0)']^{-1}), \end{aligned}$$

Lagrangian Multiplier (LM) Test

where we have used the fact that

$$\begin{aligned} A_0 I(\theta_0) A_0' &= [G(\theta_0) H(\theta_0)^{-1} G(\theta_0)']^{-1} G(\theta_0) H(\theta_0)^{-1} I(\theta_0) H(\theta_0)^{-1} G(\theta_0)' [G(\theta_0) H(\theta_0)^{-1} G(\theta_0)']^{-1} \\ &= -[G(\theta_0) H(\theta_0)^{-1} G(\theta_0)']^{-1} \end{aligned}$$

given the information matrix equality $I(\theta_0) + H(\theta_0) = \mathbf{0}$. It follows that the quadratic form

$$-n \tilde{\lambda}' G(\theta_0) H(\theta_0)^{-1} G(\theta_0)' \tilde{\lambda} \xrightarrow{d} \chi_q^2$$

under \mathbb{H}_0 .

Lagrangian Multiplier (LM) Test

Theorem 4 (9.4) [LM Test]

Suppose Assumptions M.1–M.6, Assumption W.3, and \mathbb{H}_0 hold.

Define

$$LM = n\tilde{\lambda}'G(\tilde{\theta}) \left[-\hat{H}(\tilde{\theta}) \right]^{-1} G(\tilde{\theta})'\tilde{\lambda}.$$

Then under \mathbb{H}_0 ,

$$LM \xrightarrow{d} \chi_q^2 \text{ as } n \rightarrow \infty.$$

Lagrangian Multiplier (LM) Test

Remarks:

- An asymptotically size α LM test will reject the null hypothesis $\mathbb{H}_0 : g(\theta_0) = 0$ when LM exceeds the $(1-\alpha)$ th quantile of the χ_q^2 distribution.



Question: What is the interpretation for the Lagrangian multiplier $\tilde{\lambda}$?

Lagrangian Multiplier (LM) Test

- When the parametric restriction $g(\theta_0) = 0$ is valid, the restricted estimator $\tilde{\theta}$ should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be close to zero at the restricted estimator $\tilde{\theta}$.
- As can be seen from the FOC that $G(\tilde{\theta})' \tilde{\lambda} = -\frac{d\hat{l}(\tilde{\theta})}{d\theta}$, the LM test is based on the slope of the log-likelihood at the point where the function is maximized subject to the constraint.
- Alternatively, the Lagrange multiplier $\tilde{\lambda}$ measures the magnitude of the departure of $g(\tilde{\theta})$ from $g(\theta_0) = 0$:

Lagrangian Multiplier (LM) Test

- If the constraint $g(\theta_0) = 0$ holds, then $g(\tilde{\theta})$ will be close to 0, and so is $\tilde{\lambda}$.
- If the constraint $g(\theta_0) \neq 0$, then $g(\tilde{\theta})$ will be significantly different from 0. As a result, $\tilde{\lambda}$ will be large, giving the LM test power to reject \mathbb{H}_0 . How large $\sqrt{n}\tilde{\lambda}$ is considered as “large” is determined by its sampling distribution.
- The LM test is convenient to use in practice, because only the null model, which is usually simpler, has to be estimated.

CONTENTS

9.1 Introduction to Hypothesis Testing

9.2 Neyman-Pearson Lemma

9.3 Wald Test

9.4 Lagrangian Multiplier (LM) Test

9.5 Likelihood Ratio Test

9.6 Illustrative Examples

9.7 Conclusion

Likelihood Ratio Test

- Since \mathbf{X}^n is an IID random sample from the population $f_X(x) = f(x, \theta_0)$, where θ_0 is unknown, we have the likelihood function of \mathbf{X}^n

$$f_{\mathbf{X}^n}(\mathbf{X}^n, \theta) = \prod_{i=1}^n f(X_i, \theta).$$

Likelihood Ratio Test

- We define the likelihood ratio statistic

$$\begin{aligned}\hat{\Lambda} &= \frac{\max_{\theta \in \Theta} f_{\mathbf{X}^n}(\mathbf{X}^n, \theta)}{\max_{\theta \in \Theta_0} f_{\mathbf{X}^n}(\mathbf{X}^n, \theta)} \\ &= \frac{\prod_{i=1}^n f(X_i, \hat{\theta})}{\prod_{i=1}^n f(X_i, \tilde{\theta})},\end{aligned}$$

where $\hat{\theta}$ and $\tilde{\theta}$ are the unconstrained and constrained MLEs respectively, namely,

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta \in \Theta} \hat{l}(\theta), \\ \tilde{\theta} &= \arg \max_{\theta \in \Theta_0} \hat{l}(\theta),\end{aligned}$$

with

$$\hat{l}(\theta) = \frac{1}{n} \sum_{i=1}^n \ln f(X_i, \theta)$$

is the sample average of log-likelihood functions, and Θ_0 is the parameter space Θ subject to the constraint $g(\theta) = \mathbf{0}$, i.e., $\Theta_0 = \{\theta \in \Theta : g(\theta) = \mathbf{0}\}$.

Likelihood Ratio Test

- Suppose the null hypothesis $\mathbb{H}_0 : g(\theta_0) = 0$ holds. Then both unconstrained and constrained MLEs $\hat{\theta}$ and $\tilde{\theta}$ are consistent for θ_0 , and imposing the restriction should not lead to a large reduction in the log-likelihood function. Therefore, we expect that the likelihood ratio $\hat{\Lambda}$ will be close to unity.
- On the other hand, if \mathbb{H}_0 is false, then the unconstrained MLE $\hat{\theta}$ is consistent for θ_0 but the constrained MLE $\tilde{\theta}$ is not. As a consequence, we expect that the likelihood ratio $\hat{\Lambda}$ is larger than unity.

Likelihood Ratio Test

- Hence, we can test \mathbb{H}_0 by comparing whether $\hat{\Lambda}$ is significantly larger than unity or whether $\ln \hat{\Lambda}$ is greater than 0. How large $\hat{\Lambda}$ or $\ln \hat{\Lambda}$ must be in order to be considered as significantly large will be determined by the sampling distribution of $\hat{\Lambda}$.
- Formally, we define the likelihood ratio test statistic as follows:

$$LR = 2 \ln \hat{\Lambda} = 2n \left[\hat{l}(\hat{\theta}) - \hat{l}(\tilde{\theta}) \right].$$

Likelihood Ratio Test

- By a second order Taylor's series expansion of $\hat{l}(\tilde{\theta})$ around the unconstrained MLE $\hat{\theta}$, we have

$$\begin{aligned}
 LR &= 2n \left\{ \hat{l}(\tilde{\theta}) - \left[\hat{l}(\hat{\theta}) + \frac{d\hat{l}(\hat{\theta})}{d\theta} (\tilde{\theta} - \hat{\theta}) + \frac{1}{2} (\tilde{\theta} - \hat{\theta})' \frac{d^2\hat{l}(\bar{\theta}_a)}{d\theta d\theta'} (\tilde{\theta} - \hat{\theta}) \right] \right\} \\
 &= \sqrt{n}(\tilde{\theta} - \hat{\theta})' [-\hat{H}(\bar{\theta}_a)] \sqrt{n}(\tilde{\theta} - \hat{\theta}),
 \end{aligned}$$

where $\bar{\theta}_a = a\tilde{\theta} + (1-a)\hat{\theta}$ for some $a \in [0, 1]$, lies on the segment between $\tilde{\theta}$ and $\hat{\theta}$, and $\frac{d}{d\theta}\hat{l}(\hat{\theta}) = 0$ which is the FOC of the unconstrained MLE $\hat{\theta}$, and again

$$\hat{H}(\theta) = \frac{d^2\hat{l}(\theta)}{d\theta d\theta'}$$

is the sample Hessian matrix.

Likelihood Ratio Test

- Next, applying the mean value theorem for $\frac{d}{d\theta}\hat{l}(\hat{\theta})$ around the constrained MLE $\tilde{\theta}$, we obtain

$$\begin{aligned} \mathbf{0} &= \frac{d\hat{l}(\hat{\theta})}{d\theta} \\ &= \frac{d\hat{l}(\tilde{\theta})}{d\theta} + \hat{H}(\bar{\theta}_b)(\hat{\theta} - \tilde{\theta}), \end{aligned}$$

where $\bar{\theta}_b = b\tilde{\theta} + (1-b)\hat{\theta}$ for some $b \in [0, 1]$, lies on the segment between $\tilde{\theta}$ and $\hat{\theta}$. This and the FOC of the constrained MLE that $G(\tilde{\theta})'\tilde{\lambda} = -\frac{d}{d\theta}\hat{l}(\tilde{\theta})$ imply

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \tilde{\theta}) &= -\hat{H}(\bar{\theta}_b)^{-1}\sqrt{n}\frac{d\hat{l}(\tilde{\theta})}{d\theta} \\ &= \hat{H}(\bar{\theta}_b)^{-1}G(\tilde{\theta})'\sqrt{n}\tilde{\lambda}. \end{aligned}$$

This relationship provides an alternative interpretation for the multiplier $\tilde{\lambda}$, namely, it measures the difference between the unconstrained and constrained MLEs $\hat{\theta}$ and $\tilde{\theta}$.

Likelihood Ratio Test

- It follows that

$$LR = -\sqrt{n}\tilde{\lambda}G(\tilde{\theta})\hat{H}(\bar{\theta}_b)^{-1}\hat{H}(\bar{\theta}_a)\hat{H}(\bar{\theta}_b)^{-1}G(\tilde{\theta})'\sqrt{n}\tilde{\lambda}.$$

Because

$$\begin{aligned} LR - LM &= -\sqrt{n}\tilde{\lambda} \left[G(\tilde{\theta})\hat{H}(\bar{\theta}_b)^{-1}\hat{H}(\bar{\theta}_a)\hat{H}(\bar{\theta}_b)^{-1}G(\tilde{\theta})' - G(\tilde{\theta})\hat{H}(\tilde{\theta})^{-1}G(\tilde{\theta})' \right] \sqrt{n}\tilde{\lambda} \\ &= O_P(1)o_P(1)O_P(1) \\ &= o_P(1), \end{aligned}$$

where $\sqrt{n}\tilde{\lambda} = O_P(1)$ by Lemma 7.11 given $\sqrt{n}\tilde{\lambda} \xrightarrow{d} N(0, -[G(\theta_0)H(\theta_0)^{-1}G(\theta_0)']^{-1})$ as shown in Section 9.4, and

$$\begin{aligned} &G(\tilde{\theta})\hat{H}(\bar{\theta}_b)^{-1}\hat{H}(\bar{\theta}_a)\hat{H}(\bar{\theta}_b)^{-1}G(\tilde{\theta})' - G(\tilde{\theta})\hat{H}(\tilde{\theta})^{-1}G(\tilde{\theta})' \\ \xrightarrow{p} &G(\theta_0)H(\theta_0)^{-1}H(\theta_0)H(\theta_0)^{-1}G(\theta_0)' - G(\theta_0)H(\theta_0)^{-1}G(\theta_0)' = \mathbf{0}. \end{aligned}$$

Likelihood Ratio Test

- It follows that LR and LM are asymptotically equivalent under \mathbb{H}_0 . As a result, by the asymptotic equivalence lemma and $LM \xrightarrow{d} \chi_q^2$ as $n \rightarrow \infty$ under \mathbb{H}_0 , we have the following result:

Theorem 5 (9.5) [LR Test]

Suppose Assumptions M.1–M.6, Assumption W.3, and \mathbb{H}_0 hold. Then under \mathbb{H}_0 ,

$$LR \xrightarrow{d} \chi_q^2 \text{ as } n \rightarrow \infty.$$

Likelihood Ratio Test

Remarks:

- In fact, it can also be shown that the Wald test statistic W based on MLE $\hat{\theta}$ and the LM test statistic LM are asymptotically equivalent under \mathbb{H}_0 .
- This implies that all three tests are asymptotically equivalent under \mathbb{H}_0 . The three classical tests can be graphically represented in Figure 9.x below: (Note: For the Figure, see Figure 4.8 of William Greece).
- The LR test statistic LR involves both constrained and unconstrained MLE estimators. However, it is very convenient to compute, because the sample log-likelihood value is the objective function and is usually reported by statistical software when a probability distribution model $f(x, \theta)$ is estimated by the MLE.

Likelihood Ratio Test

Theorem 6 (9.6) [LR Test Based on Sufficient Statistic]

If $T(\mathbf{X}^n)$ is a sufficient statistic for θ , and $LR(\mathbf{X}^n)$ and $LR[T(\mathbf{X}^n)]$ are the likelihood ratio tests based on \mathbf{X}^n and $T(\mathbf{X}^n)$ respectively, then

$$LR(\mathbf{X}^n) = LR[T(\mathbf{X}^n)].$$

Likelihood Ratio Test

Proof:

- By the Factorization theorem, the PMF/PDF of \mathbf{X}^n can be written as

$$f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = g[T(\mathbf{x}^n), \theta]h(\mathbf{x}^n), \text{ for all } \theta \in \Theta,$$

where $g(t, \theta)$ is the PMF/PDF of $T(\mathbf{X}^n)$ and $h(\mathbf{x}^n)$ does not depend on θ .

- It follows that

$$\begin{aligned} LR(\mathbf{X}^n) &= 2n \ln \hat{\Lambda} \\ &= 2n \ln \left[\frac{f_{\mathbf{X}^n}(\mathbf{X}^n, \hat{\theta})}{f_{\mathbf{X}^n}(\mathbf{X}^n, \tilde{\theta})} \right] \\ &= 2n \ln \left\{ \frac{g[T(\mathbf{X}^n), \hat{\theta}]h(\mathbf{X}^n)}{g[T(\mathbf{X}^n), \tilde{\theta}]h(\mathbf{X}^n)} \right\} \\ &= 2n \ln \left\{ \frac{g[T(\mathbf{X}^n), \hat{\theta}]}{g[T(\mathbf{X}^n), \tilde{\theta}]} \right\} \\ &= LR[T(\mathbf{X}^n)]. \end{aligned}$$

- This implies that the LR test statistic depends on \mathbf{X}^n only through the statistic $T(\mathbf{X}^n)$ when $T(\mathbf{X}^n)$ is a sufficient statistic for θ .

CONTENTS

9.1 Introduction to Hypothesis Testing

9.2 Neyman-Pearson Lemma

9.3 Wald Test

9.4 Lagrangian Multiplier (LM) Test

9.5 Likelihood Ratio Test

9.6 Illustrative Examples

9.7 Conclusion

Hypothesis Testing under the Bernoulli Distribution

- Suppose \mathbf{X}^n is an IID random sample from a Bernoulli(θ) distribution, where a Bernoulli random variable takes two possible values:

$$X_i = \begin{cases} 1, & \text{with probability } \theta, \\ 0, & \text{with probability } 1 - \theta. \end{cases}$$

The parameter $\theta \in (0, 1)$ is unknown.

- Suppose we are interested in testing

$$\mathbb{H}_0 : \theta = \theta_0$$

versus

$$\mathbb{H}_1 : \theta \neq \theta_0.$$

Hence, we have $g(\theta) = \theta - \theta_0$, and the gradient

$$G(\theta) = \frac{dg(\theta)}{d\theta} = 1.$$

Hypothesis Testing under the Bernoulli Distribution

- Since the population PMF $f(x, \theta) = \theta^x(1 - \theta)^{1-x}$ for $x = 0, 1$, the log-likelihood function of the IID random sample \mathbf{X}^n is given by

$$\begin{aligned}\ln \hat{L}(\theta | \mathbf{X}^n) &= \sum_{i=1}^n \ln f(X_i, \theta) \\ &= n\bar{X}_n \ln \theta + n(1 - \bar{X}_n) \ln(1 - \theta),\end{aligned}$$

where the sample mean \bar{X}_n is a sufficient statistic for θ . The FOC of the MLE is

$$\frac{\partial \ln \hat{L}(\theta | \mathbf{X}^n)}{\partial \theta} = \frac{n\bar{X}_n}{\hat{\theta}} - \frac{n - n\bar{X}_n}{1 - \hat{\theta}} = 0.$$

Thus we have the MLE $\hat{\theta} = \bar{X}_n$.

Hypothesis Testing under the Bernoulli Distribution

The Wald Test

- Recall the Hessian matrix

$$H(\theta) = E_{\theta} \left[\frac{\partial^2 \ln f(X_i, \theta)}{\partial \theta^2} \right].$$

Because

$$\frac{\partial^2 \ln f(X_i, \theta)}{\partial \theta^2} = -\frac{X_i}{\theta^2} - \frac{1 - X_i}{(1 - \theta)^2},$$

the sample Hessian matrix

$$\hat{H}(\theta) = n^{-1} \sum_{i=1}^n \frac{\partial^2 \ln f(X_i, \theta)}{\partial \theta^2} = -\frac{\sum_{i=1}^n X_i}{n\theta^2} - \frac{\sum_{i=1}^n (1 - X_i)}{n(1 - \theta)^2} = -\frac{\bar{X}_n}{\theta^2} - \frac{1 - \bar{X}_n}{(1 - \theta)^2}.$$

Hence, we have

$$\hat{H}(\hat{\theta}) = -\frac{1}{\bar{X}_n (1 - \bar{X}_n)}.$$

Hypothesis Testing under the Bernoulli Distribution

- The Wald test statistic

$$\begin{aligned}\hat{W} &= ng(\hat{\theta})' \left[-G(\hat{\theta})\hat{H}^{-1}(\hat{\theta})G(\hat{\theta})' \right]^{-1} g(\hat{\theta}) \\ &= \frac{n \left(\hat{\theta} - \theta_0 \right)^2}{\bar{X}_n (1 - \bar{X}_n)} \\ &= \frac{n \left(\bar{X}_n - \theta_0 \right)^2}{\bar{X}_n (1 - \bar{X}_n)} \xrightarrow{d} \chi_1^2 \text{ as } n \rightarrow \infty\end{aligned}$$

under \mathbb{H}_0 . Hence, $\sqrt{n}(\bar{X}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$.

Hypothesis Testing under the Bernoulli Distribution

The LM Test

- Define the Lagrangian function

$$L(\theta, \lambda) = \hat{l}(\theta) + \lambda'g(\theta) = \hat{l}(\theta) + \lambda(\theta - \theta_0)$$

where the normalized log-likelihood

$$\hat{l}(\theta) = \bar{X}_n \ln \theta + (1 - \bar{X}_n) \ln (1 - \theta).$$

- The first order conditions for the constrained MLE are

$$\begin{aligned} \frac{\partial L(\tilde{\theta}, \tilde{\lambda})}{\partial \theta} &= \frac{\partial \hat{l}(\tilde{\theta})}{\partial \theta} + \tilde{\lambda} = 0, \\ \frac{\partial L(\tilde{\theta}, \tilde{\lambda})}{\partial \lambda} &= g(\tilde{\theta}) = \tilde{\theta} - \theta_0 = 0. \end{aligned}$$

Hypothesis Testing under the Bernoulli Distribution

- It follows that $\tilde{\theta} = \theta_0$, and

$$\begin{aligned}\tilde{\lambda} &= -\frac{\partial \hat{l}(\tilde{\theta})}{\partial \theta} \\ &= -\frac{\bar{X}_n}{\tilde{\theta}} + \frac{1 - \bar{X}_n}{1 - \tilde{\theta}} \\ &= -\frac{\bar{X}_n - \tilde{\theta}}{\tilde{\theta}(1 - \tilde{\theta})} = -\frac{\bar{X}_n - \theta_0}{\theta_0(1 - \theta_0)}.\end{aligned}$$

This indicates that $\tilde{\lambda}$ measures the difference between the unconstrained MLE $\hat{\theta}$ and the constrained MLE $\tilde{\theta} = \theta_0$.

- Also, the sample Hessian matrix

$$\hat{H}(\tilde{\theta}) = -\frac{\bar{X}_n}{\theta_0^2} - \frac{1 - \bar{X}_n}{(1 - \theta_0)^2} = -\frac{\bar{X}_n(1 - \theta_0)^2 + (1 - \bar{X}_n)\theta_0^2}{\theta_0^2(1 - \theta_0)^2}$$

Hypothesis Testing under the Bernoulli Distribution

- Therefore, we have

$$\begin{aligned}
 LM &= -n\tilde{\lambda}'_n G(\tilde{\theta}) \hat{H}(\tilde{\theta})^{-1} G(\tilde{\theta})' \tilde{\lambda} \\
 &= n \left[-\frac{\bar{X}_n - \theta_0}{\theta_0(1 - \theta_0)} \right]^2 \left[\frac{\bar{X}_n(1 - \theta_0)^2 + (1 - \bar{X}_n)\theta_0^2}{\theta_0^2(1 - \theta_0)^2} \right]^{-1} \\
 &= \frac{n(\bar{X}_n - \theta_0)^2}{\bar{X}_n(1 - \theta_0)^2 + (1 - \bar{X}_n)\theta_0^2}.
 \end{aligned}$$

Hypothesis Testing under the Bernoulli Distribution

The LR Test

- Finally, we calculate the LR test statistic

$$\begin{aligned} LR &= 2n \left[\hat{l}(\hat{\theta}) - \hat{l}(\tilde{\theta}) \right] \\ &= 2n \left[\bar{X}_n \ln \left(\frac{\bar{X}_n}{\theta_0} \right) + (1 - \bar{X}_n) \ln \left(\frac{1 - \bar{X}_n}{1 - \theta_0} \right) \right]. \end{aligned}$$

Hypothesis Testing under the Normal Distribution

- Suppose X^n is an IID random sample from a $N(\mu, \sigma^2)$ population, where $\theta = (\mu, \sigma^2)$ is unknown.
- We are interested in testing the hypotheses

$$\mathbb{H}_0 : \mu = \mu_0$$

versus

$$\mathbb{H}_A : \mu \neq \mu_0,$$

where μ_0 is a known number. This is equivalent to choosing the test function

$$g(\theta) = \mu - \mu_0.$$

It follows that

$$G(\theta) = \frac{dg(\theta)}{d\theta} = (1, 0)$$

is a two-dimensional row vector.

Hypothesis Testing under the Normal Distribution

- Since the PDF of a $N(\mu, \sigma^2)$ population is

$$f(x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The normalized log-likelihood function of the random sample \mathbf{X}^n is

$$\hat{l}(\theta|\mathbf{X}^n) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

For the unconstrained MLE, we have obtained $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$, where

$$\begin{aligned}\hat{\mu} &= \bar{X}_n, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.\end{aligned}$$

Hypothesis Testing under the Normal Distribution

Also, the sample Hessian matrix

$$\hat{H}(\theta) = \begin{bmatrix} -\frac{1}{\sigma^2} & -\frac{1}{\sigma^4} \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \\ -\frac{1}{\sigma^4} \frac{1}{n} \sum_{i=1}^n (X_i - \mu) & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6} \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \end{bmatrix}.$$

When $\theta = \hat{\theta}$, we have

$$\hat{H}(\hat{\theta}) = \begin{bmatrix} -\frac{1}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{1}{2\hat{\sigma}^4} \end{bmatrix}.$$

Hypothesis Testing under the Normal Distribution

The Wald Test Statistic

$$\begin{aligned}\hat{W} &= -ng(\hat{\theta})' \left[G(\hat{\theta})\hat{H}^{-1}(\hat{\theta})G(\hat{\theta})' \right]^{-1} g(\hat{\theta}) \\ &= \frac{n(\bar{X}_n - \mu_0)^2}{\hat{\sigma}^2} \\ &\xrightarrow{d} \chi_1^2\end{aligned}$$

under \mathbb{H}_0 .

Hypothesis Testing under the Normal Distribution

The LM Test Statistic

- Consider the constrained MLE

$$\tilde{\theta} = \max_{\theta \in \Theta} \hat{l}(\theta | \mathbf{X}^n)$$

subject to the constrain that $\mu = \mu_0$.

- Define the Lagrangian function

$$L(\theta, \lambda) = \hat{l}(\theta | \mathbf{X}^n) + \lambda(\mu - \mu_0).$$

Hypothesis Testing under the Normal Distribution

- The FOCs are:

$$\frac{\partial L(\tilde{\theta}, \tilde{\lambda})}{\partial \mu} = \frac{1}{\tilde{\sigma}^2} \frac{1}{n} \sum_{i=1}^n (X_i - \tilde{\mu}) + \tilde{\lambda} = 0,$$

$$\frac{\partial L(\tilde{\theta}, \tilde{\lambda})}{\partial \sigma^2} = -\frac{1}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4} \frac{1}{n} \sum_{i=1}^n (X_i - \tilde{\mu})^2 = 0,$$

$$\frac{\partial L(\tilde{\theta}, \tilde{\lambda})}{\partial \lambda} = \tilde{\mu} - \mu_0 = 0.$$

Solving for these FOCs, we obtain

$$\begin{aligned} \tilde{\mu} &= \mu_0, \\ \tilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2, \\ \tilde{\lambda} &= -\frac{1}{\tilde{\sigma}^2} (\bar{X}_n - \mu_0), \end{aligned}$$

and the sample Hessian matrix

$$\hat{H}(\tilde{\theta}) = \begin{bmatrix} -\frac{1}{\tilde{\sigma}^2} & -\frac{1}{\tilde{\sigma}^4} (\bar{X}_n - \mu_0) \\ -\frac{1}{\tilde{\sigma}^4} (\bar{X}_n - \mu_0) & -\frac{1}{2\tilde{\sigma}^4} \end{bmatrix}.$$

Hypothesis Testing under the Normal Distribution

- It follows that the LM test statistic

$$\begin{aligned} LM &= -n\tilde{\lambda}'G(\tilde{\theta})\hat{H}^{-1}(\tilde{\theta})G(\tilde{\theta})'\tilde{\lambda} \\ &= \frac{n(\bar{X}_n - \mu_0)^2}{\tilde{\sigma}^2 - 2(\bar{X}_n - \mu_0)^2}. \end{aligned}$$

Hypothesis Testing under the Normal Distribution

The Likelihood Ratio Test Statistic

- Since

$$\begin{aligned}\hat{l}(\hat{\theta}) &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\hat{\sigma}^2) - \frac{1}{2}, \\ \hat{l}(\tilde{\theta}) &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\tilde{\sigma}^2) - \frac{1}{2}.\end{aligned}$$

- It follows that

$$\begin{aligned}LR &= 2n \left[\hat{l}(\hat{\theta}|\mathbf{X}^n) - \hat{l}(\tilde{\theta}|\mathbf{X}^n) \right] \\ &= n \ln \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right).\end{aligned}$$

- It is interesting to observe that the LR principle tests the hypothesis on the population mean by comparing the variance estimators under the null and alternative hypotheses. Here, the likelihood ratio is a log-function of the sample variance ratio. Intuitively, when $\mu \neq \mu_0$, $\tilde{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \mu_0)^2$ is not a consistent estimator for σ^2 . It will be larger than $\hat{\sigma}^2$ when n is large, giving the LR its power.

CONTENTS

9.1 Introduction to Hypothesis Testing

9.2 Neyman-Pearson Lemma

9.3 Wald Test

9.4 Lagrangian Multiplier (LM) Test

9.5 Likelihood Ratio Test

9.6 Illustrative Examples

9.7 Conclusion

Conclusion

- Hypothesis testing is one of the most important tasks in statistical inference.
- In this chapter we have introduced basic ideas of hypothesis testing in statistical inference.
- We introduce the well-known Neyman-Pearman lemma that a likelihood ratio based test will be uniformly most powerful test for simple hypotheses.
- We discuss three important testing methods, namely the Wald test, the Lagrange Multiplier (LM) test, and the Likelihood Ratio (LR) test, and show they are asymptotically equivalent to each other under the null hypothesis.

Conclusion

- It is important to note that all hypothesis tests considered in this chapter assume that the population distribution model is correctly specified.
- When the population distribution model is misspecified, the Wald test statistic and the LM test statistics have to be modified to use a consistent asymptotic variance estimator which is robust to model misspecification. However, it is impossible to modify the Likelihood ratio test statistic.
- Finally, when testing economic hypotheses, we usually need to transform an economic hypothesis into a statistical hypothesis on model parameters. Since some auxiliary conditions are often imposed in such a transformation, there usually exists a gap between the original economic hypothesis and the resulting statistical hypothesis. Caution is needed to interpret the testing results of the statistical hypotheses.