Module 6: Statistical inference (III)

Siyue Yang

06/01/2022

Outline

This module we will review

- Basics of hypothesis testing
- P-values
- The Wald test
- The score test
- The likelihood ratio test

Hypothesis testing

Definition (Hypothesis testing)

Suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

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

We call H_0 the null hypothesis and H_1 the alternative hypothesis.

Rejection region

Let X be a random variable and let $\mathcal{X}$ be the range of X. Rejection region is a subset of outcomes $R \in \mathcal{X}$

$$
\begin{aligned}
X \in R \quad &\Longrightarrow \quad \text{reject } H_0 \\
X \notin R \quad &\Longrightarrow \quad \text{retain (do not reject) } H_0
\end{aligned}
$$

Usually, the rejection region is

$$R = \{x : T(x) > c\}$$

where T is a test statistic and c is a critical value.

Type I error and type II error

	Retain Null	Reject Null
H_0 true	$\surd$	type I error
H_1 true	type II error	$\surd$

Power function and the size of a test

Definition (Power function)

The power function of a test with rejection region R is defined by

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

Definition (The size of a test)

The size of a test is defined to be

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

A test is said to have level α if its size is less than or equal to α.

Exercise

Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ where σ is known. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Hence, $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$.

Consider the test:

$$\text{reject } H_0 \text{ if } T > c$$

where $T = \bar{X}$. The rejection region is

$$R = \{(x_1, \ldots, x_n) : T(x_1, \ldots, x_n) > c\}$$

What is the power function? What is the size of the test?

Exercise (cont'd)

Let Z denote a standard Normal random variable. The power function is

$$\begin{aligned}\beta(\mu) &= \mathbb{P}_\mu(\bar{X} > c) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} > \frac{\sqrt{n}(c-\mu)}{\sigma}\right) \\ &= \mathbb{P}\left(Z > \frac{\sqrt{n}(c-\mu)}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right)\end{aligned}$$

Exercise (cont'd)

$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right)$$

For a size α test, we set this equal to α and solve for c to get

$$c = \frac{\sigma \Phi^{-1}(1-\alpha)}{\sqrt{n}}$$

We reject when $\bar{X} > \sigma \Phi^{-1}(1-\alpha)/\sqrt{n}$. Equivalently, we reject when

$$\frac{\sqrt{n}(\bar{X} - 0)}{\sigma} > z_\alpha$$

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

Most powerful test

The test with highest power under H_1, among all size α tests (if it exists), is called **most powerful**.

In the special case of a simple null $H_0 : \theta = \theta_0$ and a simple alternative $H_1 : \theta = \theta_1$ we can say precisely what the most powerful test is.

Most powerful test

The test with highest power under H_1, among all size α tests (if it exists), is called **most powerful**.

In the special case of a simple null $H_0 : \theta = \theta_0$ and a simple alternative $H_1 : \theta = \theta_1$ we can say precisely what the most powerful test is.

Definition (Neyman-Pearson Lemma)

Suppose we test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Let

$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i=1}^{n} f(x_i; \theta_1)}{\prod_{i=1}^{n} f(x_i; \theta_0)}$$

Suppose we reject H_0 when $T > k$. If we choose k so that $\mathbb{P}_{\theta_0}(T > k) = \alpha$ then this test is the most powerful, size α test. That is, among all tests with size α, this test maximizes the power $\beta(\theta_1)$.

P-values

Definition (P-values)

Suppose that for every $\alpha \in (0,1)$ we have a size α test with rejection region R_α. Then,

$$\text{p-value } = \inf \left\{ \alpha : T\left(X^n\right) \in R_\alpha \right\}.$$

That is, the p-value is the smallest level at which we can reject H_0.

Misconceptions of P-value

- A large p-value is not strong evidence in favor of H_0. A large p-value can occur for two reasons: (i) H_0 is true or (ii) H_0 is false but the test has low power.

Misconceptions of P-value

- A large p-value is not strong evidence in favor of H_0. A large p-value can occur for two reasons: (i) H_0 is true or (ii) H_0 is false but the test has low power.
- The p-value is not the probability that the null hypothesis is true.

P-values (cont'd)

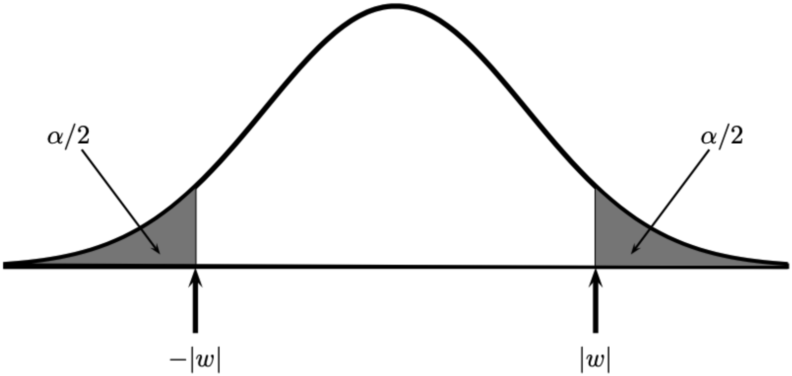


FIGURE 10.4. The p-value is the smallest α at which you would reject H_0. To find the p-value for the Wald test, we find α such that $|w|$ and $-|w|$ are just at the boundary of the rejection region. Here, w is the observed value of the Wald statistic: $w = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$. This implies that the p-value is the tail area $\mathbb{P}(|Z| > |w|)$ where $Z \sim N(0, 1)$.

Widely used tests

1. Wald test
2. Score test
3. Likelihood ratio test

The Wald test

Consider testing

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

using log-likelihood function $\ell(\theta)$.

Intuitively, the farther $\widehat{\theta}_n$ is from θ_0, the stronger the evidence against the null hypothesis.

How far is "far enough"?

The Wald test (cont'd)

We use the fact that under regularity assumptions that we have under H_0,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, I^{-1}\left(\theta_0\right)\right)$$

where

$$I\left(\theta_0\right) = \mathbb{E}_{\theta_0}\left[\frac{\partial^2 \log f(X \mid \theta)}{\partial \theta^2}\right]$$

- Wald statistics:

$$W_n = \sqrt{nI\left(\theta_0\right)}\left(\widehat{\theta}_n - \theta_0\right)$$

The Wald test (cont'd)

Under H_0,

$$W_n = \sqrt{n\hat{I}(\theta_0)}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$$

- Rejects H_0 if $|W_n| \geq z_{\alpha/2}$, where $P\left(Z \geq z_{\alpha/2}\right) = \alpha/2$.
- Asymptotic size α test

$$\mathbb{P}_{\theta_0}\left(|W_n| > z_{\alpha/2}\right) \to \mathbb{P}_{\theta_0}\left(|Z| > z_{\alpha/2}\right) = \alpha$$

Power

- Suppose the true value of θ is $\theta_\star \neq \theta_0$. The power $\beta\left(\theta_\star\right) -$ the probability of correctly rejecting the null hypothesis - is given (approximately) *by*

$$1 - \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathrm{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathrm{se}}} - z_{\alpha/2}\right)$$

Power

- Suppose the true value of θ is $\theta_\star \neq \theta_0$. The power $\beta\left(\theta_\star\right)-$ the probability of correctly rejecting the null hypothesis - is given (approximately) *by*

$$1-\Phi\left(\frac{\theta_0-\theta_\star}{\widehat{\mathrm{se}}}+z_{\alpha/2}\right)+\Phi\left(\frac{\theta_0-\theta_\star}{\widehat{\mathrm{se}}}-z_{\alpha/2}\right)$$

- If $\theta_\star$ far from θ_0, or the sample size is large, power is large.

Size of the Wald test

- The size α Wald test rejects $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where

$$C = \left(\widehat{\theta} - \widehat{\text{se}} z_{\alpha/2}, \hat{\theta} + \widehat{\text{se}} z_{\alpha/2}\right)$$

Size of the Wald test

- The size α Wald test rejects $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where

$$C = \left(\widehat{\theta} - \widehat{\mathsf{se}} z_{\alpha/2}, \hat{\theta} + \widehat{\mathsf{se}} z_{\alpha/2}\right)$$

- Testing the hypothesis is equivalent to checking whether the null value is in the confidence interval.

Statistically significant v.s. Scientifically significant

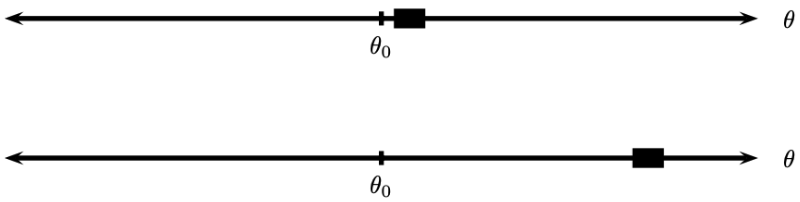


FIGURE 10.2. Scientific significance versus statistical significance. A level α test rejects $H_0 : \theta = \theta_0$ if and only if the $1 - \alpha$ confidence interval does not include θ_0. Here are two different confidence intervals. Both exclude θ_0 so in both cases the test would reject H_0. But in the first case, the estimated value of θ is close to θ_0 so the finding is probably of little scientific or practical value. In the second case, the estimated value of θ is far from θ_0 so the finding is of scientific value. This shows two things. First, statistical significance does not imply that a finding is of scientific importance. Second, confidence intervals are often more informative than tests.

Beyond MLE estimate

- Wald test is not limited to MLE estimate, you just need to know the asymptotic distribution of your test satistic.
- Example: Assume we have $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from populations with mean μ_1 and v.
- We write $\delta = \mu_1 - \mu_2$ and we want to test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$.
- We build

$$W = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{m}}}$$

  where S_1^2 and S_2^2 are the sample variances.
- Thanks to the *CLT*, we have $W \xrightarrow{\mathrm{D}} \mathcal{N}(0, 1)$ as $m, n \to \infty$.

The score test

Under $H_0 : \theta = \theta_0$

$$\frac{1}{\sqrt{n}} \ell' (\theta_0) \xrightarrow{\mathrm{D}} \mathcal{N}(0, I(\theta_0))$$

where

$$\ell'(\theta) = \frac{\partial \log L(\theta \mid \mathbf{x})}{\partial \theta}$$

- Score statistic

$$R_n = \frac{\ell' (\theta_0)}{\sqrt{nI(\theta_0)}}$$

Proof sketch (Optional)

$$0 = \ell'\left(\widehat{\theta}_n\right) \approx \ell'(\theta_0) + \ell''(\theta_0)\left(\widehat{\theta}_n - \theta_0\right)$$

thus

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) \approx -\frac{\ell''(\theta_0)}{\sqrt{n}}\left(\widehat{\theta}_n - \theta_0\right) = -\frac{\ell''(\theta_0)}{n}\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$$

where

$$-\frac{\ell''(\theta_0)}{n} \xrightarrow{\mathrm{P}} I(\theta_0) \text{ and } \sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, I^{-1}(\theta_0)\right)$$

The result follows from Slutzky's lemma.

The score test (cont'd)

Under H_0,

$$R_n = \frac{\ell'(\theta_0)}{\sqrt{nI(\theta_0)}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$$

- Rejects H_0 if $|R_n| \geq z_{\alpha/2}$, where $P\left(Z \geq z_{\alpha/2}\right) = \alpha/2$

The likelihood ratio test

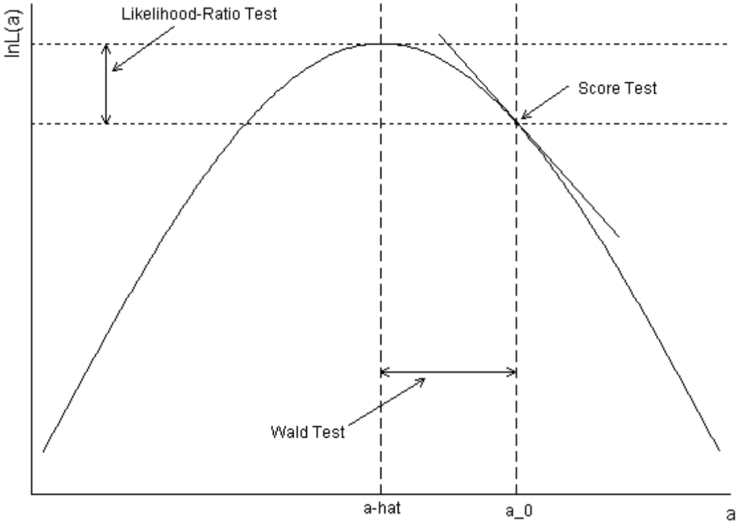
$$\Delta_n = l\left(\widehat{\theta}_n\right) - l\left(\theta_0\right) = \log\left(\frac{\sup_{\theta \in \Theta}(\theta \mid \mathbf{x})}{L\left(\theta_0 \mid \mathbf{x}\right)}\right) \geq 0$$

Under H_0,

$$2\Delta_n \xrightarrow{\mathrm{D}} \chi_1^2$$

- As the $1-\alpha$ quantile of a χ_1^2 distribution is $z_{\alpha/2}^2$,
- we rejects H_0 when $2\Delta_n \geq z_{\alpha/2}^2$.
- i.e. We reject small values of LR test statistics.

The Wald test, score test, and likelihood ratio test



Fox, J. (1997) Applied regression analysis, linear models, and related methods. Thousand Oaks, CA: Sage Publications. P. 570.

Test equivalence

We can show that (when there is no misspecification)

$$R_n \xrightarrow{\mathrm{P}} W_n$$
$$W_n^2 \xrightarrow{\mathrm{P}} 2\Delta_n.$$

- The tests are thus asymptotically equivalent in the sense that under H_0 they reach the same decision with probability 1 as $n \to \infty$.
- For a finite sample size n, they have some relative advantages and disadvantages with respect to one another.

Discussion

$$W_n = \sqrt{n\hat{I}(\theta_0)}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$$

$$R_n = \frac{\ell'(\theta_0)}{\sqrt{nI(\theta_0)}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$$

$$2\Delta_n = 2\left\{l\left(\widehat{\theta}_n\right) - l(\theta_0)\right\} \xrightarrow{\mathrm{D}} \chi_1^2$$

- It is easy to create one-sided Wald and score tests.
- The score test does not require $\widehat{\theta}_n$ whereas the other two tests do.
- The Wald test is most easily interpretable and yields immediate confidence intervals.
- The score test and LR test are invariant under reparametrization, whereas the Wald test is not.

Resources

This tutorial is based on

- “All of statistics” Chapter 10 by Larry A. Wasserman.
- Arnaud Doucet’s STA 461 Lecture notes [links].