

**Single Parameter Case: inference based on asymptotic normality**

In the previous lecture we discussed an example where we had a random sample from a normal distribution with unknown mean  $\mu$  and known variance  $\sigma^2$ , and we wished to test the hypothesis

$$H_0 : \mu = \mu_0,$$

against the alternative

$$H_1 : \mu \neq \mu_0.$$

We proposed a test at the 5% significance level, which rejected the null if

$$\left| \frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu_0) \right| > 1.96.$$

Equivalently, we could work with the test statistic

$$\frac{n}{\sigma^2} (\bar{X}_n - \mu_0)^2,$$

which is distributed  $\chi^2(1)$  under the null hypothesis, and reject the null if this statistic is greater than 3.84.

The key to doing this, was that  $\bar{X}_n$  had a known distribution under  $H_0$ . However, in many other settings, the exact distribution of the test statistic is not easily calculable, and finding exact critical regions is practically infeasible. One way to deal with this problem, similar to what we did in point estimation, is to use large sample approximations.

Suppose that  $X_i, i = 1, \dots, n$  are IID with PDF/PMF  $f_X(x; \theta)$ , and to start with, suppose that the parameter  $\theta$  is a scalar. We want to test the null hypothesis that  $\theta = \theta_0$  against the alternative that  $\theta \neq \theta_0$ . By our previous results, we know that the MLE  $\hat{\theta}$  is asymptotically normal:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1}),$$

where  $I(\theta) = E[(\frac{\partial}{\partial \theta} \log f_X(X; \theta))^2]$  is the single-observation Fisher information.

If the null hypothesis is true, then the above display holds with  $\theta = \theta_0$ . If we replace  $I(\theta_0)$  with a consistent estimator  $\hat{I} \xrightarrow{p} I(\theta_0)$ , then by Slutsky's lemma,

$$\sqrt{n\hat{I}}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, 1).$$

In other words, the statistic  $\sqrt{n\hat{I}}(\hat{\theta} - \theta_0)$  is approximately standard normal under the null, and

$$Pr \left( \left| \sqrt{n\hat{I}}(\hat{\theta} - \theta_0) \right| > 1.96 \right) \rightarrow 0.05.$$

So a test that rejects  $H_0 : \theta = \theta_0$  if the  $\left| \sqrt{n\hat{I}}(\hat{\theta} - \theta_0) \right| > 1.96$  will be have significance level approximately 0.05. If we want to use a different significance level, such as .10, we could adjust the critical value appropriately.

If we want to construct an approximate confidence interval for  $\theta$ , we can use a similar line of reasoning. Now, we need that

$$\hat{I} \xrightarrow{P} I(\theta)$$

whatever is the true value of  $\theta$ . To simplify the notation, let

$$SE = \left(\sqrt{n\hat{I}}\right)^{-1}.$$

This is often called a “standard error,” and it is an estimator of the standard deviation of  $\hat{\theta}$ . Then, by the definition of convergence in distribution,

$$\begin{aligned} Pr\left(-1.96 \leq \frac{\hat{\theta} - \theta}{SE} \leq 1.96\right) &\rightarrow 0.95 \\ \Rightarrow Pr\left(-1.96 \leq \frac{\theta - \hat{\theta}}{SE} \leq 1.96\right) &\rightarrow 0.95 \\ \Rightarrow Pr\left(\hat{\theta} - 1.96SE \leq \theta \leq \hat{\theta} + 1.96SE\right) &\rightarrow 0.95. \end{aligned}$$

Thus, the interval  $\hat{\theta} \pm 1.96SE$  is an approximate 95% confidence interval for  $\theta$ .

Notice that our argument really just relies on having an asymptotically normal point estimator and some estimate of its standard deviation. So we could also use a similar line of reasoning for other asymptotically normal estimators. It is common to report point estimates along with standard errors when reporting the results of a statistical analysis. From these two pieces of information, you can reconstruct a hypothesis test or a confidence interval fairly easily.

We do need to come up with an estimator  $\hat{I}$ , or equivalently, the standard error. In some parametric models you can derive a relatively simple expression for the Fisher information and use this to come up with a simple estimator. In more complicated models you can replace the expectation

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \log f_X(X; \theta)\right)^2\right]$$

with its sample average, using the estimate  $\hat{\theta}$  in place of the (unknown) true value of  $\theta$ :

$$\hat{I} = \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial}{\partial\theta} \log f_X(X_i; \hat{\theta})\right)^2.$$

### Multiple Parameter Case

Suppose now that  $\theta$  is a  $k$ -dimensional vector. Then the MLE is asymptotically multivariate normal

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N_k(0, I(\theta)^{-1}),$$

where  $\hat{\theta}$  and  $\theta_0$  are vectors and  $I(\theta)$  is the  $k \times k$  information matrix

$$I(\theta) = E \left[ \left( \frac{\partial}{\partial \theta} \log f_X(X_i; \theta) \right) \left( \frac{\partial}{\partial \theta} \log f_X(X_i; \theta) \right)' \right].$$

This implies that for any particular element of  $\theta$ , say  $\theta_j$ , we have:

$$\sqrt{n}(\hat{\theta}_j - \theta_j) \xrightarrow{d} N_1(0, V_j),$$

where

$$V_j = [I(\theta)^{-1}]_{jj},$$

is the  $(j, j)$  element of the variance-covariance matrix  $I(\theta)^{-1}$ . As before, we replace  $I(\theta)$  by an estimator  $\hat{I}$  in the expression for  $V_j$  to construct standard errors for each element of  $\hat{\theta}_j$ . If we want to test a hypothesis about a particular element of  $\theta_j$ , or construct a confidence interval for  $\theta_j$ , we can then repeat the analysis of the previous section.

### Single Parameter Case: Wald, LR, and LM tests

Returning to the case where  $\theta$  is a scalar, and consider the problem of testing  $H_0 : \theta = \theta_0$  against the alternative  $H_a : \theta \neq \theta_0$ . Under the null hypothesis,

$$\sqrt{n}\hat{I}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, 1).$$

We can square this quantity to obtain the Wald statistic

$$WALD = n\hat{I}(\hat{\theta} - \theta_0)^2.$$

By the continuous mapping theorem,

$$WALD \xrightarrow{d} \chi^2(1).$$

Let  $C_\alpha$  be the value such that  $Pr(\chi^2(1) > C_\alpha) = \alpha$ . For example, if  $\alpha = .05$ , then  $C_\alpha = 3.84$ , and if  $\alpha = .10$ , then  $C_\alpha = 2.706$ . (You can get these critical values from a standard Table of the Chi-Square distribution.) Then

$$Pr(WALD > C_\alpha) \rightarrow \alpha.$$

In other words, the Type I error probability of the Wald test is converging to the desired significance level.

There are two other tests for the same null and alternative hypothesis that give very similar results in large samples, and in fact are large sample equivalent. Indirectly all three tests are based on the likelihood function. If the null hypothesis is correct, the log likelihood function should be close to the expected log likelihood function, which is maximized at  $\theta_0$ . Therefore its maximum should be close to the  $\theta_0$  (the Wald test), the maximizing value should be close to the value at  $\theta_0$  (the likelihood ratio test), and the derivative at  $\theta_0$  should be close to zero (the Lagrange multiplier test).

The likelihood ratio test is based on the maximum value of the log likelihood function under the null and under the alternative hypothesis. Define

$$\lambda = \frac{\max_{\theta \in \Theta_0^c} f_X(x; \theta)}{\max_{\theta \in \Theta_0} f_X(x; \theta)}.$$

In our case the null hypothesis consists of a single point  $\theta_0$ , so the denominator is just  $f_X(x; \theta_0)$ . Also, in the numerator, the value that maximizes the likelihood under the alternative is the maximum likelihood estimator (except if the maximum likelihood estimator is exactly equal to  $\theta_0$ , but that is very unlikely). Therefore:

$$\lambda = \frac{f_X(x; \hat{\theta})}{f_X(x; \theta_0)}.$$

In addition we have  $n$  independent and identically distributed random variables so,

$$\lambda = \frac{\mathcal{L}(x_1, \dots, x_n; \hat{\theta})}{\mathcal{L}(x_1, \dots, x_n; \theta_0)}.$$

Under the null hypothesis,

$$LR = 2 \cdot \log \lambda \xrightarrow{d} \mathcal{X}^2(1),$$

a chi-squared distribution with one degree of freedom. So we can use the same “cutoffs” as in the Wald test.

To see the connection with the Wald test, expand  $L(\theta_0)$  around  $\hat{\theta}$ :

$$\begin{aligned} L(\theta_0) &= L(\hat{\theta}) + \frac{\partial L}{\partial \theta}(\hat{\theta}) \cdot (\theta_0 - \hat{\theta}) + \frac{1}{2} \frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) \cdot (\theta_0 - \hat{\theta})^2 \\ &= L(\hat{\theta}) + \frac{1}{2} \frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) \cdot (\theta_0 - \hat{\theta})^2, \end{aligned}$$

for some  $\tilde{\theta}$  between  $\hat{\theta}$  and  $\theta_0$ . Note that  $\frac{\partial L}{\partial \theta}(\hat{\theta}) = 0$ . Substituting this into  $\lambda$  gives

$$2 \cdot \log \lambda = 2 \cdot \left( L(\hat{\theta}) - L(\theta_0) \right) = -\frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) \cdot (\theta_0 - \hat{\theta})^2.$$

Note that

$$\frac{1}{N} \frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) \xrightarrow{p} -I(\theta_0).$$

Hence

$$2 \cdot \log \lambda \rightarrow N \cdot (\theta_0 - \hat{\theta})^2 \cdot I(\theta_0) \approx \text{WALD}.$$

The Lagrange multiplier or score test, the third test in the trinity of tests works off the fact that the expectation of the score function is zero: if the null hypothesis is true, then

$$E \left[ \frac{\partial}{\partial \theta} \log f_X(X; \theta_0) \right] = 0.$$

The variance of the score is

$$I(\theta_0) = E \left[ \left( \frac{\partial}{\partial \theta} \log f_X(X; \theta_0) \right)^2 \right].$$

Hence, using a central limit theorem,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_X(X_i; \theta_0) \xrightarrow{d} N(0, I(\theta_0)).$$

Given that we do not know the exact information matrix, we use an estimate. Typically we use the test statistic

$$LM = \frac{1}{n} \left( \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_X(x_i; \theta_0) \right)^2 / \hat{I}(\theta_0).$$

To see that this is again similar to the Wald test expand the sum of the derivative of the log of the density around  $\hat{\theta}$ :

$$\begin{aligned} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_X(x_i; \theta_0) &= \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_X(x_i; \hat{\theta}_0) + \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_X(x_i; \tilde{\theta}) \cdot (\theta_0 - \hat{\theta}) \\ &= \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_X(x_i; \tilde{\theta}) \cdot (\theta_0 - \hat{\theta}) \\ &\approx -N \cdot I(\theta_0) \cdot (\theta_0 - \hat{\theta}), \end{aligned}$$

so that

$$\begin{aligned} LM &= \frac{1}{n} \left( \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_X(x_i; \theta_0) \right)^2 / \hat{I}(\theta_0) \approx \frac{1}{n} \left( -n \cdot I(\theta_0) \cdot (\theta_0 - \hat{\theta}) \right)^2 / \hat{I}(\theta_0) \\ &\approx n \cdot I(\theta_0) \cdot (\hat{\theta} - \theta_0)^2 \approx WALD. \end{aligned}$$

### Example

Let us consider an example. Suppose  $X_1, \dots, X_n$  are IID with a Poisson distribution with mean  $\theta$ . Its PMF is

$$f_X(x; \theta) = \frac{\theta^x \exp(-\theta)}{x!}.$$

We wish to test the null hypothesis

$$H_0 : \theta = 6,$$

against the alternative hypothesis

$$H_a : \theta \neq 6,$$

at the 10% level. It is given that  $n = 100$  and  $\sum_{i=1}^n x_i = 500$ .

First consider the maximum likelihood estimator. The log likelihood function is

$$L(\theta) = \sum_{i=1}^n x_i \cdot \log \theta - \theta - \log x_i!,$$

and the maximum likelihood estimate is  $\hat{\theta} = \bar{x} = 5$ . We also have

$$\frac{\partial}{\partial \theta} \log f_X(x; \theta) = \frac{x}{\theta} - 1;$$

$$\frac{\partial^2}{\partial \theta^2} \log f_X(x; \theta) = -\frac{x}{\theta^2}.$$

The single-observation information matrix is therefore

$$I(\theta) = V(X)/\theta^2 = \theta/\theta^2 = 1/\theta,$$

using the square of the first derivatives, or

$$I(\theta) = E[X]/\theta^2 = \theta/\theta^2 = 1/\theta,$$

using the second derivatives. The information matrix is estimated by evaluating it at the maximum likelihood estimate,

$$\hat{I}_1 = 1/\hat{\theta} = 1/5,$$

or, for tests of the hypothesis that  $\theta = 6$ , we can use

$$\hat{I}_2 = 1/\theta_0 = 1/6.$$

To carry out a normal-based test using  $\hat{I}_1$ , we can calculate the statistic

$$\sqrt{n\hat{I}_1}(\hat{\theta} - \theta_0) = 10\sqrt{\frac{1}{5}}(5 - 6) = -4.47.$$

This is greater in absolute value than the 10% cutoff for the normal distribution, which is 1.645, so we reject the null hypothesis. (Note: we could have used  $\hat{I}_2$  instead here; this would give slightly different results but the same conclusion.) We can also form confidence intervals by calculating the standard error

$$SE = \left( \sqrt{n\hat{I}_1} \right)^{-1} = 0.224.$$

Next, consider the LR, Wald, and LM tests. The value of the log likelihood function under the null is

$$\begin{aligned} L(\theta_0) &= \sum_{i=1}^n x_i \cdot \log \theta_0 - \theta_0 - \log x_i! \\ &= 500 \cdot \log 6 - 100 \cdot 6 - \sum_{i=1}^n \log x_i!. \end{aligned}$$

The value of the log likelihood function at the maximum likelihood estimator is

$$\begin{aligned} L(\hat{\theta}) &= \sum_{i=1}^n x_i \cdot \log \hat{\theta} - \hat{\theta} - \log x_i! \\ &= 500 \cdot \log 5 - 100 \cdot 5 - \sum_{i=1}^n \log x_i!. \end{aligned}$$

Twice the difference is

$$LR = 2 \cdot \left( L(\hat{\theta}) - L(\theta_0) \right) = 2 \cdot \left( 500 \log 5 - 500 - 500 \log 6 + 600 \right) \approx 17.6$$

Next, consider the Wald test:

$$WALD = N \cdot (\theta_{ml} - \theta_0)^2 \cdot \hat{I}(\hat{\theta}) = 100 \cdot (6 - 5)^2 \cdot 1/6 = 16.7.$$

Finally, consider the Lagrange multiplier test. The sum of the derivatives is

$$\sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_X(x_i; \theta_0) = \sum_{i=1}^n \frac{x}{\theta_0} - 1 = 500/6 - 100 = -16.7.$$

The Lagrange multiplier test statistic is therefore

$$LM = \frac{1}{n} \left( \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_X(x_i; \theta_0) \right)^2 / \hat{I}(\theta_0) = \frac{1}{100} \cdot (-16.67)^2 / (1/6) \approx 16.7.$$

Note that with the Lagrange multiplier test we typically use the estimator for the information matrix based on the value of the parameter under the null hypothesis, as a key benefit of the Lagrange multiplier test is that it does not require estimation of the model under the alternative hypothesis.

In all cases the test statistic exceeds the critical value for a  $\chi^2(1)$  distribution at the 10% level, which is 2.706. Typically the test statistics are close enough that the result (rejection or acceptance) does not depend on the actual test chosen, although this does happen occasionally.  $\square$

If we are willing to make large sample approximations, then we have three basic tests that are approximately equivalent, and have different practical advantages and disadvantages. The Wald test is easy to calculate once the maximum likelihood estimator and its large sample variance have been calculated. The likelihood ratio test has the advantage of not requiring an arbitrary choice for the information matrix. The Lagrange multiplier has the advantage of not requiring estimation of the model under the alternative hypothesis.