

(T3) Slutsky's Theorem: $X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X) \quad \forall$ continuous Functions $g(\cdot)$

Combining (T2) w/ Slutsky's Theorem we have the following Corollaries:

(C3a) Corollary: $X_n \xrightarrow{P} x$ and $Y_n \xrightarrow{P} y \Rightarrow$
(a) $X_n + Y_n \xrightarrow{P} x + y$
(b) $X_n Y_n \xrightarrow{P} xy$
(c) $X_n / Y_n \xrightarrow{P} x / y$ if $y \neq 0$ a.s.

(C3b) Corollary: $A_n \xrightarrow{P} A$ and $B_n \xrightarrow{P} B \Rightarrow$
(a) $A_n + B_n \xrightarrow{P} A + B$
(b) $A_n B_n \xrightarrow{P} AB$
(c) $A_n^{-1} \xrightarrow{P} A^{-1}$ if A^{-1} exists a.s.

Remark: The reason we need (T2) to apply (T3) to (C3a) and (C3b) is that addition, multiplication and division/inversion are continuous functions of (X_n, Y_n) and (A_n, B_n) . Thus, we need joint convergence to apply Slutsky's Theorem. (T2) gives us this joint convergence from the individual convergence assumptions. The same is not true, in general, for convergence in distribution. That is, in general, $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y \not\Rightarrow (X_n, Y_n) \xrightarrow{d} (X, Y)$. We do, however, have the following result:

(T4) Theorem: $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} Y_0 \Rightarrow (X_n, Y_n) \xrightarrow{d} (X, Y_0)$

Note: Recall that $Y_n \xrightarrow{P} Y_0 \Leftrightarrow Y_n \xrightarrow{d} Y_0$

(T5) Continuous Mapping Theorem: $X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X) \quad \forall$ continuous Functions $g(\cdot)$

(C5a) Corollary: $X_n \xrightarrow{d} x$ and $Y_n \xrightarrow{d} y_0 \Rightarrow$
(a) $X_n + Y_n \xrightarrow{d} x + y_0$ ($\Rightarrow Y_n \xrightarrow{P} y_0$)
(b) $X_n Y_n \xrightarrow{d} x y_0$
(c) $X_n / Y_n \xrightarrow{d} x / y_0$ if $y_0 \neq 0$

(C5b) Corollary: $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} Y_0$ and $A_n \xrightarrow{P} A_0 \Rightarrow$
 (a) $X_n + Y_n \xrightarrow{d} X + Y_0$
 (b) $A_n X_n \xrightarrow{d} A_0 X$

(T6) Theorem: $X_n \xrightarrow{d} X$ and $X_n - Y_n \xrightarrow{P} 0 \Rightarrow Y_n \xrightarrow{d} X$

LAWS OF LARGE NUMBERS

(T7) Markov's Inequality: $X \geq 0 \Rightarrow P\{X \geq \epsilon\} \leq EX / \epsilon \quad \forall \epsilon > 0$

(T8) Chebyshev's Inequality: $P\{\|X - Y\| \geq \epsilon\} \leq E\|X - Y\|^2 / \epsilon^2 \quad \forall$ random vectors X, Y and $\forall \epsilon > 0, \epsilon > 0$.

(T9) Jensen's Inequality: $Eg(X) \geq g(EX)$ if $g(\cdot)$ is convex
 $\Rightarrow Eg(X) \leq g(EX)$ if $g(\cdot)$ is concave.

(T10) WLLN: $\{X_i\}_n$ iid random vectors w/ $E\|X_i\|^2 < \infty \Rightarrow \bar{X}_n \xrightarrow{P} EX_i$

(T11) Khintchine's SLLN: $\{X_i\}_n$ iid random vectors w/ $E\|X_i\| < \infty \Rightarrow \bar{X}_n \xrightarrow{a.s.} EX_i$

(T12) Markov's LLN: $\{X_i\}_n$ independent random vectors such that $\exists \delta > 0$ for which $\frac{1}{n} \sum_{i=1}^n E\|X_i\|^{1+\delta}$ bounded $\forall n \Rightarrow \bar{X}_n - EX_n \xrightarrow{P} 0$

(T13) Chebyshev's LLN: $\{X_i\}_n$ stationary random vectors w/ $E\|Y\|^2 < \infty$ and $\sum_{j=1}^{\infty} \|Cov(X_i, X_{i+j})\| < \infty \Rightarrow \bar{X}_n \xrightarrow{P} EX_i$

Note: A process is said to be strictly stationary if $\forall i$ and $\forall j_1, \dots, j_m$ the joint distribution of $(X_i, X_{i+j_1}, \dots, X_{i+j_m})$ depends only on j_1, \dots, j_m and not i . It is said to be weakly stationary or Covariance Stationary if $EX_i = \mu \quad \forall i$ and $Cov(X_i, X_{i+j}) = \Sigma_j \quad \forall i, j$. The last LLN requires only weak stationarity.

(T10) and (T12) require iid
 (T12) requires independence

(T13) requires stationarity but not independence.

CENTRAL LIMIT THEOREMS

(T14) Lindberg-Levy CLT: X_i iid random vectors w/ $E X_i = \mu$ and $V(X_i) = \Sigma$

$\Rightarrow \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma)$

(T15) Lindberg-Feller CLT: X_i are independent random vectors such that

- (i) $E X_i = \mu_i$ Finite $\forall i$ define $\bar{\mu}_n = \frac{1}{n} \sum_{i=1}^n \mu_i$
- (ii) $V(X_i) = \Sigma_i$ Finite $\forall i$ define $\bar{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n \Sigma_i$
- (iii) All mixed third moments of X_i are Finite $\forall i$
- (iv) $\bar{Q}_n \rightarrow Q$ Finite PD
- (v) $(n \bar{Q}_n)^{-1} Q_i \rightarrow 0 \forall i$

$\bar{Q}_n = \bar{\Sigma}_n$

$\Rightarrow \sqrt{n}(\bar{X}_n - \bar{\mu}_n) \xrightarrow{d} N(0, Q)$

(T16) Delta-Method: IF Z_n is a sequence of $k \times 1$ vectors such that $\sqrt{n}(Z_n - z_0) \xrightarrow{d} N(0, \Sigma)$ for some $k \times 1$ vector z_0 and $g: \mathbb{R}^k \rightarrow \mathbb{R}^J$

$(J \times k)$ is continuously differentiable at z_0 , then

$\sqrt{n}(g(Z_n) - g(z_0)) \xrightarrow{d} N(0, G \Sigma G')$

where $G = \partial g(z_0) / \partial z'$ is the $J \times k$ matrix of partial derivatives at z_0

Proof:

$\sqrt{n}(g(Z_n) - g(z_0)) = \frac{\partial g(\bar{z})}{\partial z'} \sqrt{n}(Z_n - z_0)$ For some \bar{z} between Z_n and z_0 by the

Mean Value Theorem. $\sqrt{n}(Z_n - z_0) \xrightarrow{d} N(0, \Sigma) \Rightarrow Z_n \xrightarrow{p} z_0 \Rightarrow \bar{z} \xrightarrow{p} z_0$. Thus, by (T3) $\frac{\partial g(\bar{z})}{\partial z'} \xrightarrow{p} G$ and by (C5b) (b) we get the desired result.

(2) MAXIMUM LIKELIHOOD ESTIMATION

Consider a sample of iid random vectors Z_1, \dots, Z_n from a joint pdf $f(Z_i | \theta)$ where the unknown parameter vector θ_0 is the object of interest to be estimated. Then we have

Likelihood Function: $L(\theta | Z) = \prod_{i=1}^n f(Z_i | \theta)$

Log-likelihood Function: $\ell(\theta | Z) = \ln L(\theta | Z) = \sum_{i=1}^n \ln f(Z_i | \theta)$

where we will usually write $L(\theta)$ and $\ell(\theta)$ instead of $L(\theta|Z)$ and $\ell(\theta|Z)$. The Maximum Likelihood Estimator (MLE) of θ_0 is:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} L(\theta) = \underset{\theta}{\operatorname{argmax}} \ell(\theta)$$

We also define the Information Matrix of a sample of size n as:

$$I_n(\theta_0) = -E \left[\frac{\partial^2 \ell(\theta_0)}{\partial \theta' \partial \theta} \right] = E \left[\left(\frac{\partial \ell(\theta_0)}{\partial \theta} \right) \left(\frac{\partial \ell(\theta_0)}{\partial \theta} \right)' \right] = V \left[\frac{\partial \ell(\theta_0)}{\partial \theta} \right]$$

where the second and third equalities were proven in 14.381. Since the data are iid it also follows that the Information Matrix of a single observation, denoted $I_1(\theta_0)$ satisfies the relation:

$$I_1(\theta_0) = \frac{1}{n} I_n(\theta_0)$$

$$I_n = -E \left[\frac{\partial^2 \ell}{\partial \theta' \partial \theta} \right] = -E \left[\frac{\partial^2 (\sum \ell_i)}{\partial \theta' \partial \theta} \right] = \sum I_{i,n}$$

We now have the following theorems:

(T17) Theorem: (Under regularity conditions) IF $\hat{\theta}$ is any unbiased estimator of θ_0 , then $V(\hat{\theta}) - [I_n(\theta_0)]^{-1}$ is PSD. $[I_n(\theta_0)]^{-1}$ is the Cramér-Rao Lower Bound, written $CRLB = [I_n(\theta_0)]^{-1}$.

(T18) Theorem: (Under regularity conditions) we have

(a) $\hat{\theta}_{ML} \xrightarrow{P} \theta_0$

(b) $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, [I_1(\theta_0)]^{-1})$

(c) $\hat{\theta}_{ML}$ is asymptotically efficient

(d) The MLE of $\gamma = g(\theta_0)$ is $\hat{\gamma} = g(\hat{\theta}_{ML})$

(e) IF $\hat{\theta}_{ML}$ is unbiased, it achieves the CRLB and so is BUE

Also

(f) $\frac{1}{n} \frac{\partial^2 \ell(\theta_0)}{\partial \theta' \partial \theta} \xrightarrow{d} N(0, I_1(\theta_0))$

$$\left. \begin{aligned} (g) \frac{1}{n} \frac{\partial^2 \ell(\bar{\theta})}{\partial \theta' \partial \theta} &= -\frac{1}{n} \sum \left[\frac{\partial^2 \ln f(z; \bar{\theta})}{\partial \theta' \partial \theta} \right] \xrightarrow{P} I_1(\theta_0) \\ (h) \frac{1}{n} \sum \begin{bmatrix} \frac{\partial \ln f(z; \bar{\theta})}{\partial \theta} & \frac{\partial \ln f(z; \bar{\theta})}{\partial \theta'} \end{bmatrix} &\xrightarrow{P} I_1(\theta_0) \end{aligned} \right\} \begin{array}{l} \forall \bar{\theta} \text{ s.t.} \\ \bar{\theta} \xrightarrow{P} \theta_0 \end{array}$$

Note: An estimator is asymptotically efficient if it is asymptotically normal (i.e. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$) and its asymptotic variance matrix V is "smaller" than that of any other asymptotically normal estimator, where by smaller we mean that the difference is PSD.

It is essential to note here that in general, the finite sample properties of $\hat{\theta}_{MLE}$ are unknown. In particular, (T18) does not imply that $\hat{\theta}_{MLE}$ is unbiased or normally distributed in small samples (although it may be in some cases).

(3) Hypothesis Testing

Whenever you want to test a null hypothesis, H_0 , against some alternative hypothesis, H_1 , you use a test-statistic based on the data (which means that it is a random variable). The test procedure consists of deciding whether or not you accept H_0 by comparing the realization of the test-statistic to some preset number called the critical value of the test-statistic. More generally, you examine whether or not the realization of the test statistic falls inside some critical region (e.g. values \geq the critical value). If so, you reject H_0 in favor of H_1 , otherwise you accept H_0 (or do not reject H_0). Usually, H_0 represents some restrictions placed on H_1 and so is nested inside H_1 . Your test is therefore a test of the restrictions imposed by H_0 . A testing procedure can give a wrong result in two ways:

Type I Error: H_0 , although true, is rejected $\rightarrow \alpha$

Type II Error: H_0 , although false, is accepted $\rightarrow \beta$

Some Definitions:

- SIZE of the test $\equiv \alpha = \overset{\text{maximum}}{\text{probability of a Type I error}}$. This is also called the significance level of the test. $= \text{Prob}\{\text{reject } H_0 \mid H_0 \text{ is true}\}$
- power of the test $\equiv 1 - \beta = 1 - \text{probability of a Type II error} = 1 - \beta$
 $= \text{Prob}\{\text{reject } H_0 \mid H_0 \text{ is false}\}$

$$P(\theta) = \text{Prob}\{X \in C_\alpha^* | \theta\}$$

critical region

The power of a test will generally vary over Θ_2 , the alternative parameter space

- A test that has greater power (smaller β) than all other tests of the same size (same α) is called most powerful (for a given value of the parameter to be tested)
- A test which is most powerful for all possible values of the underlying parameter is called uniformly most powerful
- A test is unbiased if its power is greater than its size (ie $1 - \beta \geq \alpha$) for all values of the underlying parameter
 $\text{Prob}\{reject H_0 | \theta_0\} \rightarrow \text{Prob}\{reject H_0 | true\}$
 $\text{Prob}\{X \in C | \theta\} > \text{Prob}\{X \in C | \theta_0\}$
- A test is consistent if its power goes to 1 as the sample size goes to infinity.

(E1) Ex: Two-Sided Test of the Mean of a Normal Distribution

$X_i \sim N(\mu, \sigma^2)$ iid μ, σ^2 unknown

$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ← estimator of μ $\bar{X} \sim N(\mu, \sigma^2/n)$

$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ ← estimator of σ^2 $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ ind. of \bar{X}

$H_0: \mu = \mu_0$ vs. $H_2: \mu \neq \mu_0$

$T = \sqrt{n}(\bar{X} - \mu_0)/S$ ← test-statistic, $T \sim t_{n-1}$ under H_0

Reject H_0 if $|T| \geq C_{\alpha/2}$ For a test of size α

where $P(t_{n-1} \geq C_{\alpha/2}) = \alpha/2$

(E2) EX: One-Sided Tests of the Mean of a Normal Distribution

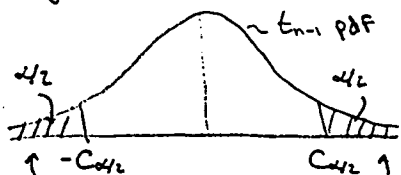
$H_0: \mu \leq \mu_0$ vs. $H_1: \mu > \mu_0$

Reject if $T \geq C_\alpha$

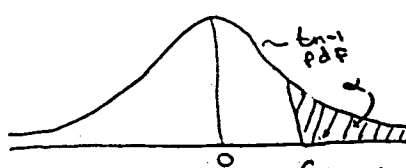
OR

$H_0: \mu \geq \mu_0$ vs. $H_2: \mu < \mu_0$

Reject if $T \leq -C_\alpha$



Two-sided Test



Critical Region



Critical Region

One Sided Tests

(E3) Ex: Two-Sided Test of the Variance of a Normal Distribution

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{vs.} \quad H_1: \sigma^2 \neq \sigma_0^2$$

$$W = (n-1)S^2/\sigma_0^2 \sim \chi^2_{n-1} \quad \text{under } H_0$$

Reject H_0 if $W \leq \underline{c}_{\alpha/2}$ or $W \geq \bar{c}_{\alpha/2}$ where $P(\chi^2_{n-1} \leq \underline{c}_{\alpha/2}) = \alpha/2$
and $P(\chi^2_{n-1} \geq \bar{c}_{\alpha/2}) = \alpha/2$

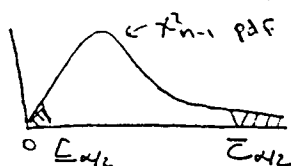
(E4) Ex: One-Sided Tests ...

$$H_0: \sigma^2 \leq \sigma_0^2 \quad \text{vs.} \quad H_1: \sigma^2 > \sigma_0^2$$

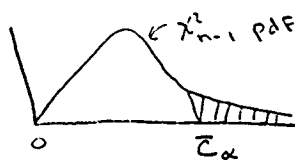
Reject H_0 if $W \geq \bar{c}_\alpha$

$$H_0: \sigma^2 \geq \sigma_0^2 \quad \text{vs.} \quad H_1: \sigma^2 < \sigma_0^2$$

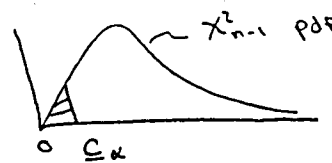
Reject H_0 if $W \leq \underline{c}_\alpha$



Two-Sided Test



One Sided Tests



(4) WALD, LR, and LM TESTS

Consider maximum likelihood estimation of an unknown parameter vector θ_0 and a test of the hypotheses:

$$H_0: h(\theta_0) = 0 \quad \text{vs.} \quad H_1: h(\theta_0) \neq 0$$

where $h(\theta_0)$ is a $J \times 1$ vector (ie imposes J restrictions on θ_0) with $\text{rank}(\partial h(\theta)/\partial \theta') = J$ (ie no redundant restrictions). Note that in order for this to be true we must have $J \leq k$ where θ_0 is $k \times 1$. Further, since $J = k$ uniquely identifies the restricted value of θ_0 , we often assume that $J < k$.

$$\hat{\theta}_u: \text{unrestricted MLE: } \hat{\theta}_u = \text{argmax } \ell(\theta)$$

$$\hat{\theta}_R: \text{restricted MLE: } \hat{\theta}_R = \text{argmax } \ell(\theta) \quad \text{s.t. } h(\theta) = 0$$

$$H(\theta) = \partial h(\theta)/\partial \theta'$$

Wald Test: The Wald Test is based on the observation that in large samples $h(\hat{\theta}_n) \approx h(\theta_0)$ which, if H_0 is true, is itself 0. Thus, the Wald Statistic for the test of $h(\theta_0) = 0$ is:

$$W = h(\hat{\theta}_n)' [H(\hat{\theta}_n) [\hat{I}_n(\hat{\theta}_n)]^{-1} H(\hat{\theta}_n)']^{-1} h(\hat{\theta}_n)$$

where $\frac{1}{n} \hat{I}_n(\hat{\theta}_n) \xrightarrow{P} I_1(\theta_0)$

Lagrange Multiplier Test: This test is based on the observation that if the restrictions imposed on θ by $h(\cdot)$ are true, then the slope of the loglikelihood function evaluated at the restricted MLE, $\hat{\theta}_R$, is approximately 0. Equivalently, the vector of Lagrange multipliers, λ , from the restricted maximization problem are approximately zero since by the FOC's:

$$\frac{\partial \ell(\hat{\theta}_R)}{\partial \theta} = H(\hat{\theta}_R)' \lambda$$

If H_0 is true, (Note: don't confuse the hypotheses H_0 and H_1 and the derivative matrix $H(\theta)$), we would expect

$$\frac{\partial \ell(\hat{\theta}_R)}{\partial \theta} \approx \frac{\partial \ell(\hat{\theta}_n)}{\partial \theta} \stackrel{\leftarrow}{=} 0 \text{ by the FOC's of unrestricted MLE}$$

The Lagrange Multiplier Statistic for testing $h(\theta_0) = 0$ is therefore

$$LM = \left[\frac{\partial \ell(\hat{\theta}_R)}{\partial \theta} \right]' [\hat{I}_n(\hat{\theta}_R)]^{-1} \left[\frac{\partial \ell(\hat{\theta}_R)}{\partial \theta} \right]$$

where $\frac{1}{n} \hat{I}_n(\hat{\theta}_R) \xrightarrow{P} I_1(\theta_0)$ under H_0 .

Likelihood Ratio Test: The likelihood ratio test is based on the observation that if the restrictions are true, then the ratio of the likelihood functions $L(\hat{\theta}_R)/L(\hat{\theta}_n) \approx 1$, otherwise we would expect $L(\hat{\theta}_R)/L(\hat{\theta}_n) \ll 1$. The Likelihood Ratio Statistic is actually based on the log of this ratio (recall $\ln(1) = 0$) and takes the form:

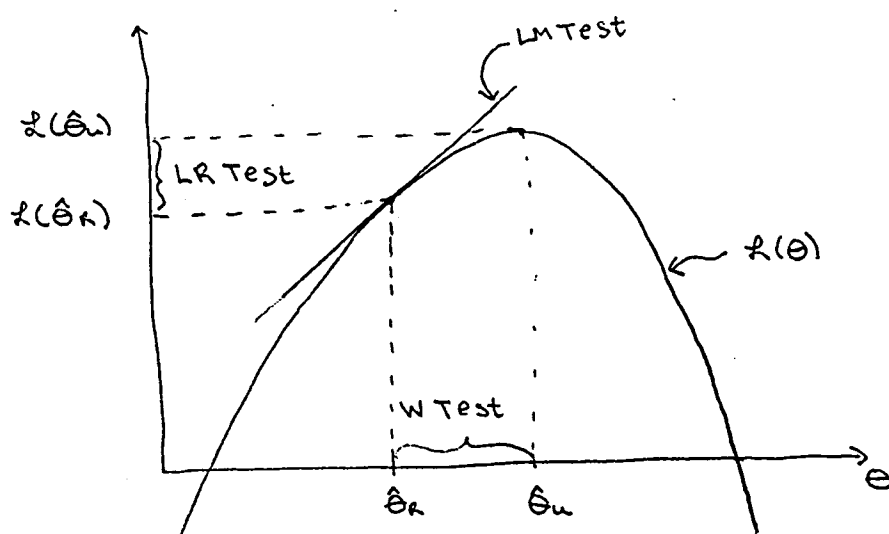
$$LR = 2[\ell(\hat{\theta}_W) - \ell(\hat{\theta}_R)]$$

(T19) Theorem: Under H_0 we have $W \xrightarrow{d} \chi^2_J$, $LM \xrightarrow{d} \chi^2_J$ and $LR \xrightarrow{d} \chi^2_J$. Thus, three asymptotically equivalent tests of H_0 vs. H_1 are given by (1) reject H_0 if $W \geq C_\alpha$ (2) reject H_0 if $LM \geq C_\alpha$ and (3) reject H_0 if $LR \geq C_\alpha$ where $P(\chi^2_J \geq C_\alpha) = \alpha$.

Remarks:

- The Wald Test requires only $\hat{\theta}_W$.
- The Lagrange Multiplier Test requires only $\hat{\theta}_R$.
- The Likelihood Ratio Test requires both $\hat{\theta}_W$ and $\hat{\theta}_R$ but is very simple to compute once you have them.
- The equivalence of these tests is asymptotic not numerical and so in practice they could yield conflicting conclusions. Indeed, in finite samples it is often the case that $W > LR > LM$ making the Wald Test the most likely to reject H_0 for a given α and the Lagrange Multiplier Test the least likely to reject. The reason for this is that the finite sample sizes of the tests are actually not n .

Finally, to get a visual sense of what each test is doing, consider the case where θ is a scalar and we are testing $H_0: \theta_0 = \bar{\theta}$ vs. $H_1: \theta_0 \neq \bar{\theta}$. In this case, $\hat{\theta}_R = \bar{\theta}$ and we have the following graph



Proof of (T19)

We begin by establishing some facts that will help us with the proof:

- (1) $\hat{\theta}_u \xrightarrow{P} \theta_0$ and, under H_0 , $\hat{\theta}_R \xrightarrow{P} \theta_0$ } by Theorem (T18)
 (a) (b) and by my
 (2) $\sqrt{n}(\hat{\theta}_u - \theta_0) \xrightarrow{d} N(0, [I_2(\theta_0)]^{-1})$ } assumption for $\hat{\theta}_R$
- (3) $\frac{\partial \ell(\hat{\theta}_u)}{\partial \theta} = 0$ by FOCs of unrestricted MLE

(4) $\frac{\partial \ell(\hat{\theta}_R)}{\partial \theta} = H(\hat{\theta}_R)' \lambda$ } by FOCs of restricted MLE

(5) $h(\hat{\theta}_R) = 0$

(6) $h(\hat{\theta}_u) = \frac{h(\hat{\theta}_R)}{0} + H(\bar{\theta})(\hat{\theta}_u - \hat{\theta}_R)$ by Mean Value Theorem for some

$\bar{\theta}$ between $\hat{\theta}_u$ and $\hat{\theta}_R$. Note: $\bar{\theta} \xrightarrow{P} \theta_0$ under $H_0 \Rightarrow H(\bar{\theta}) \xrightarrow{P} H(\theta_0)$

(7) $\frac{\partial \ell(\hat{\theta}_R)}{\partial \theta} = \frac{\partial \ell(\hat{\theta}_u)}{\partial \theta} + \frac{\partial^2 \ell(\bar{\theta})}{\partial \theta' \partial \theta} (\hat{\theta}_R - \hat{\theta}_u)$ by Mean Value Theorem

for some $\bar{\theta}$ between $\hat{\theta}_u$ and $\hat{\theta}_R$. Note $\bar{\theta} \xrightarrow{P} \theta_0$ under H_0 and so $-\frac{1}{n} \frac{\partial^2 \ell(\bar{\theta})}{\partial \theta' \partial \theta} \xrightarrow{P} I_2(\theta_0)$ under H_0 by (T18) (g).

$$\Rightarrow (\hat{\theta}_u - \hat{\theta}_R) = \left[-\frac{\partial^2 \ell(\bar{\theta})}{\partial \theta' \partial \theta} \right]^{-1} \frac{\partial \ell(\hat{\theta}_R)}{\partial \theta}$$

(8) $\ell(\hat{\theta}_R) = \ell(\hat{\theta}_u) + \frac{\partial \ell(\hat{\theta}_u)}{\partial \theta} (\hat{\theta}_R - \hat{\theta}_u) + \frac{1}{2} (\hat{\theta}_R - \hat{\theta}_u)' \frac{\partial^2 \ell(\hat{\theta}_u)}{\partial \theta' \partial \theta} (\hat{\theta}_R - \hat{\theta}_u) + o_p(1)$

$$\Rightarrow LR = (\hat{\theta}_u - \hat{\theta}_R)' \left[-\frac{\partial^2 \ell(\hat{\theta}_u)}{\partial \theta' \partial \theta} \right] (\hat{\theta}_u - \hat{\theta}_R) + o_p(1)$$

$2[\ell(\hat{\theta}_u) - \ell(\hat{\theta}_R)]$

We will now prove the theorem. The proof requires repeated applications of our rules for convergence in probability and distribution so I will not do out or explain each and every step, but the idea of the proof should be obvious from the steps I do give:

$$W \xrightarrow{d} \chi^2_3$$

$$\bullet \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, [I_2(\theta_0)]^{-1}) \quad \text{by (2)}$$

$$\Rightarrow \sqrt{n}(h(\hat{\theta}_n) - h(\theta_0)) \xrightarrow{d} N(0, \underbrace{H(\theta_0) [I_2(\theta_0)]^{-1} H(\theta_0)'}_{3 \times 3}) \quad \text{by Delta Method}$$

$$\bullet H(\hat{\theta}_n) \xrightarrow{p} H(\theta_0) \quad \text{by Slutsky's Theorem}$$

$$\Rightarrow \sqrt{n}[h(\hat{\theta}_n) - h(\theta_0)]' [H(\hat{\theta}_n) [\hat{I}_n(\hat{\theta}_n)]^{-1} H(\hat{\theta}_n)']^{-1} \sqrt{n}[h(\hat{\theta}_n) - h(\theta_0)] \xrightarrow{d} \chi^2_3$$

Under H_0 , $h(\theta_0) = 0$. Thus, noticing that the \sqrt{n} 's cancel w/ the $\frac{1}{n}$, we have

$$W = h(\hat{\theta}_n)' [H(\hat{\theta}_n) [\hat{I}_n(\hat{\theta}_n)]^{-1} H(\hat{\theta}_n)']^{-1} h(\hat{\theta}_n) \xrightarrow{d} \chi^2_3 \quad \text{under } H_0$$

$$LH \xrightarrow{d} \chi^2_3$$

Let $d\lim(X_n)$ denote the limiting distribution of X_n (kind of like $p\lim$ but w/ a "d"). Then, under H_0 , we have

$$d\lim(W) = d\lim \left\{ h(\hat{\theta}_n)' [H(\hat{\theta}_n) [\hat{I}_n(\hat{\theta}_n)]^{-1} H(\hat{\theta}_n)']^{-1} h(\hat{\theta}_n) \right\}$$

$$= d\lim \left\{ n h(\hat{\theta}_n)' [H(\theta_0) [I_2(\theta_0)]^{-1} H(\theta_0)']^{-1} h(\hat{\theta}_n) \right\} \quad \text{recall } I_2(\theta) = \frac{1}{n} \hat{I}_n(\theta)$$

$$= d\lim \left\{ n (\hat{\theta}_n - \theta_0)' H(\theta_0) [H I_2^{-1} H']^{-1} H(\theta_0) (\hat{\theta}_n - \theta_0) \right\} \quad \text{by (6) letting}$$

$$H = H(\theta_0) \quad \text{and} \quad I_2 = I_2(\theta_0)$$

$$= d\lim \left\{ n (\hat{\theta}_n - \theta_0)' H' [H I_2^{-1} H']^{-1} H (\hat{\theta}_n - \theta_0) \right\}$$

↓

$$= \text{dlim} \left\{ n \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right]' \left[-\frac{\partial^2 \lambda(\bar{\theta})}{\partial \theta' \partial \theta} \right]^{-1} H' [H \Gamma_1^{-1} H']^{-1} H \left[-\frac{\partial^2 \lambda(\bar{\theta})}{\partial \theta' \partial \theta} \right] \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\} \text{ by (7)}$$

$$= \text{dlim} \left\{ \frac{1}{n} \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right]' \left[-\frac{1}{n} \frac{\partial^2 \lambda(\bar{\theta})}{\partial \theta' \partial \theta} \right]^{-1} H' [H \Gamma_1^{-1} H']^{-1} H \left[-\frac{1}{n} \frac{\partial^2 \lambda(\bar{\theta})}{\partial \theta' \partial \theta} \right] \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\}$$

$$= \text{dlim} \left\{ \frac{1}{n} \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right]' \Gamma_1^{-1} H' [H \Gamma_1^{-1} H']^{-1} H \Gamma_1 \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\} \text{ by (118) (g)}$$

$$= \text{dlim} \left\{ \frac{1}{n} \lambda' H(\hat{\theta}_R) \Gamma_1^{-1} H' [H \Gamma_1^{-1} H']^{-1} H \Gamma_1 \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\} \text{ by (4)}$$

$$= \text{dlim} \left\{ \frac{1}{n} \lambda' \frac{H \Gamma_1^{-1} H' [H \Gamma_1^{-1} H']^{-1} H \Gamma_1}{\text{Identity Matrix}} \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\}$$

$$= \text{dlim} \left\{ (H' \lambda)' (n \Gamma_1)^{-1} \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\}$$

$$= \text{dlim} \left\{ [H(\hat{\theta}_R)' \lambda]' [I_n(\hat{\theta}_R)]^{-1} \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\}$$

$$= \text{dlim} \left\{ \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right]' [I_n(\hat{\theta}_R)]^{-1} \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\} = \overline{\text{dlim(LM)}} //$$

LR $\rightarrow \chi^2$

$$\text{dlim(LR)} = \text{dlim} \left\{ (\hat{\theta}_u - \hat{\theta}_R)' \left[-\frac{\partial^2 \lambda(\hat{\theta}_u)}{\partial \theta' \partial \theta} \right]^{-1} (\hat{\theta}_u - \hat{\theta}_R) \right\} \text{ by (8)}$$

$$= \text{dlim} \left\{ \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right]' \left[-\frac{\partial^2 \lambda(\bar{\theta})}{\partial \theta' \partial \theta} \right]^{-1} \left[-\frac{1}{n} \frac{\partial^2 \lambda(\hat{\theta}_u)}{\partial \theta' \partial \theta} \right] \left[-\frac{1}{n} \frac{\partial^2 \lambda(\bar{\theta})}{\partial \theta' \partial \theta} \right] \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\} \text{ by (7)}$$

$$= \text{dlim} \left\{ \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right]' [I_n(\hat{\theta}_R)]^{-1} I_1 \Gamma_1^{-1} \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\}$$

$$= \text{dlim} \left\{ \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right]' [I_n(\hat{\theta}_R)]^{-1} \left[\frac{\partial \lambda(\hat{\theta}_R)}{\partial \theta} \right] \right\} = \text{dlim(LM)} //$$

(5) ORDERS OF MAGNITUDE

For non stochastic sequences $\{a_n\}$ and $\{b_n\}$

- $a_n = o(b_n)$ if $a_n/b_n \rightarrow 0$ (a_n is of order smaller than b_n)
- $a_n = O(b_n)$ if a_n/b_n is asymptotically bounded (a_n is at most of order b_n)

For a sequence of random variables $\{X_n\}$ and a non stochastic sequence $\{b_n\}$

- $X_n = o_p(b_n)$ if $X_n/b_n \xrightarrow{P} 0$ (X_n is of order smaller than b_n in probability.)
- $X_n = O_p(b_n)$ if $\forall \epsilon > 0, \exists$ a constant $C(\epsilon)$ s.t. $P\{|X_n/b_n| > C(\epsilon)\} < \epsilon$
 $\forall n$ (X_n is at most of order b_n in probability)

(T20) Theorem: Some rules for stochastic orders of magnitude

$$(1) O_p(a_n) + O_p(b_n) = O_p(\max\{a_n, b_n\})$$

$$(2) O_p(a_n) O_p(b_n) = O_p(a_n b_n)$$

$$(3) O_p(a_n) + O_p(b_n) = O_p(\max\{a_n, b_n\})$$

$$(4) O_p(a_n) O_p(b_n) = O_p(a_n b_n)$$

$$(5) o_p(a_n) + o_p(b_n) = o_p(\max\{a_n, b_n\})$$

$$(6) o_p(a_n) o_p(b_n) = o_p(a_n b_n)$$

(C20) Corollary: $O_p(1) O_p(1) = O_p(1)$

Note: $X_n = o_p(1) \Leftrightarrow X_n \xrightarrow{P} 0$