# Lec 11. Likelihood Ratio Tests

Zhang W.P.

# Contents

11.1 Likelihood Ratio Test . . . . . 1
11.2 Find the LRT . . . . . 3
11.3 Asymptotic distribution of LR . . . . . 4
11.4 Wald test . . . . . 9
11.5 Score test . . . . . 11

## Learning objectives

After this section, you should be able to ...

- understand the motivation of likelihood ratio (LR) test
- understand the connections between LR test, Wald test and Score test
- find the LR tests for specific problems

## 11.1 Likelihood Ratio Test

In this section we introduce intuitive and efficient procedures that can be used when no optimal methods are available and that are natural for multi-dimensional parameters.

Suppose the sample $\mathbf{X} = (X_1, \ldots, X_n)$ has pdf/pmf $f(\mathbf{x}, \theta)$, and consider

$$H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$$

where $\Theta = \Theta_0 \cup \Theta_1$ is a subset of $\mathbb{R}^k$.

- For $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, the optimal test statistic is

$$\frac{f(\mathbf{X}, \theta_1)}{f(\mathbf{X}, \theta_0)}$$

- If we have composite hypotheses, we could instead use

$$\tilde{\Lambda}_n = \frac{\sup_{\theta \in \Theta_1} f(\mathbf{X}, \theta)}{\sup_{\theta \in \Theta_0} f(\mathbf{X}, \theta)}.$$

Since the critical value will be larger than 1, it will not change the test if we replace this statistic by $\tilde{\Lambda}_n \vee 1$.

Denote the likelihood ratio (LR)

$$\Lambda_n(\mathbf{x}) = \left(\tilde{\Lambda}_n \vee 1\right) = \frac{\left(\sup_{\theta\in\Theta_1} f(\mathbf{x},\theta)\right) \vee \left(\sup_{\theta\in\Theta_0} f(\mathbf{x},\theta)\right)}{\sup_{\theta\in\Theta_0} f(\mathbf{x},\theta)}$$

$$= \frac{\sup_{\theta\in\Theta_0\cup\Theta_1} f(\mathbf{x},\theta)}{\sup_{\theta\in\Theta_0} f(\mathbf{x},\theta)} = \frac{L(\widehat{\theta}_n)}{L(\widehat{\theta}_n^R)}$$

Def.

Then a level $\alpha$ likelihood ratio test (LRT) is given by

$$\phi(\mathbf{x}) = \begin{cases} 1, & \Lambda_n(\mathbf{x}) > c \\ r, & \Lambda_n(\mathbf{x}) = c \\ 0, & \Lambda_n(\mathbf{x}) < c \end{cases}$$

with $c, r$ satisfying $E_\theta\phi(\mathbf{X}) \leq \alpha$ for all $\theta \in \Theta_0$.

## 11.2 Find the LRT

General procedure to find the LRT of $H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$:

- Calculate the unconstrained MLE $\widehat{\theta}_n$ of $\theta$.
- Calculate the constrained MLE $\widehat{\theta}_n^R$ of $\theta$ under $\Theta_0$.
- Form $\Lambda_n(\mathbf{x})$
- **Find a function** $h$ that is strictly increasing on the range of $\Lambda_n$ such that $h(\Lambda_n(\mathbf{X}))$ has a simple form and a tabled distribution under $H_0$. Because $h(\lambda(\mathbf{X}))$ is equivalent to $\Lambda_n(\mathbf{X})$, we specify the size $\alpha$ likelihood ratio test through the test statistic $h(\Lambda_n(\mathbf{X}))$

Some examples.

## 11.3 Asymptotic distribution of LR

**Theorem 1** (Wilks). *Let $X_1, \ldots, X_n$ be a simple random sample from $f(x, \theta)$, $\theta \in \Theta$, where $\Theta$ is an open set of $\mathbb{R}^k$. Consider $H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta - \Theta_0$, where $\Theta_0$ is an open set of $\mathbb{R}^d$ with $d = k - s \geq 1$. Under regularity conditions (Conditions for MLE to be consistent and asymptotic normal) and assuming $H_0$ is true, then*

$$2log\Lambda_n(\mathbf{X}) \xrightarrow{\mathcal{L}} \chi^2_{k-d} = \chi^2_s, \quad as\ n \to \infty$$

**Note**: Wilks's theorem depends critically on the fact that not only is $\Theta$ open but that $\Theta_0$ is open.

Specification on $\Theta_0$:

- There exist a non-empty subset $A$ in $\mathbb{R}^d, d = k - s \geq 1$ and one-to-one mapping $h = (h_1 \ldots, h_k) : A \mapsto \Theta_0$ where $\partial h / \partial \theta$ is full rank of $d$ at the interior of $A$, that is,

$$\Theta_0 = \{\theta = h(\eta), \eta \in A\}$$

such that,

$$\{f(x,\theta), \theta \in \Theta_0\} = \{f(x, h(\eta)), \eta \in A\}$$

- Or, $\Theta_0$ can be equivalently expressed (after some transformation) as

$$\Theta_0 = \{\theta = (\theta_1, \ldots, \theta_k) \in \Theta : \theta_i = \theta_{i0}, i = 1, \ldots, s\}$$

and $(\theta_{10}, \ldots, \theta_{s0}, \theta_{s+1}, \ldots, \theta_{s+k})$ is interior of $\Theta$.

- Or the null can be expressed as $s$ constraints in general:

$$H_0 : g(\theta_0) = 0_{s\times 1}$$

where $g : \mathbb{R}^k \mapsto \mathbb{R}^s$ and the rank of $\partial g(\theta)/\partial\theta$ is $s$.

- In a word, $dim\Theta - dim\Theta_0 = s \geq 1$

*Proof.* Let $\theta_0$ be the true parameter, thus under $H_0$ we have $\theta_0 = h(\eta_0)$. Write the log likelihood as

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f\left(X_i, \theta\right)$$

and under $H_0$,

$$\tilde{\ell}_n(\eta) = \sum_{i=1}^{n} \log f\left(X_i, h(\eta)\right)$$

It is easy to see that the Fisher information

$$\tilde{I}(\eta) = BI(\theta)B^T, B = \left(\frac{\partial h}{\partial \eta}\right)_{d\times k}$$

Because the MLE $\widehat{\theta}_n$ satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \approx I^{-1}(\theta_0)\frac{1}{\sqrt{n}}\nabla_\theta \ell_n(\theta_0) \equiv I^{-1}(\theta_0)U_n(\theta_0),$$

where $\nabla_\theta$ is the derivative with respect to $\theta$ and $U_n(\theta) = \frac{1}{\sqrt{n}}\nabla\ell_n(\theta)$. An expansion of $\ell_n(\theta)$ about $\widehat{\theta}_n$ evaluated at $\theta = \theta_0$ gives

$$\begin{aligned} 2\left[\ell_n(\widehat{\theta}_n) - \ell_n\left(\theta_0\right)\right] &\approx n\left(\widehat{\theta}_n - \theta_0\right)^T I\left(\theta_0\right)\left(\widehat{\theta}_n - \theta_0\right) \\ &\approx U_n(\theta_0)^T I^{-1}(\theta_0) U_n(\theta_0), \end{aligned}$$

Similar argument applied to $\tilde{\ell}_n(\eta)$ to get

$$\sqrt{n}(\widehat{\eta}_n - \eta_0) \approx \tilde{I}^{-1}(\eta_0)\frac{1}{\sqrt{n}}\nabla_\eta\tilde{\ell}_n(\eta_0) \equiv \tilde{I}^{-1}(\eta_0)\tilde{U}_n(\eta_0),$$

and

$$\begin{aligned} 2\left[\tilde{\ell}_n(\widehat{\eta}_n) - \tilde{\ell}_n\left(\eta_0\right)\right] &\approx n\left(\widehat{\eta}_n - \eta_0\right)^T \tilde{I}\left(\eta_0\right)\left(\widehat{\eta}_n - \eta_0\right) \\ &\approx \tilde{U}_n(\eta_0)^T \tilde{I}^{-1}(\eta_0)\tilde{U}_n(\eta_0) \\ &= U_n(\theta_0)^T B^T \tilde{I}^{-1}(\eta_0) B U_n(\theta_0), \end{aligned}$$

where $\tilde{U}_n(\eta) = \frac{1}{\sqrt{n}}\nabla\tilde{\ell}_n(\eta) = BU_n(\theta)$.

Therefor,

$$\begin{aligned} 2log\Lambda_n(\mathbf{X}) &= 2\left[\ell_n(\widehat{\theta}_n) - \ell_n\left(\theta_0\right)\right] - 2\left[\tilde{\ell}_n(\widehat{\eta}_n) - \tilde{\ell}_n\left(\eta_0\right)\right] \\ &\approx U_n(\theta_0)^T[I^{-1}(\theta_0) - B^T\tilde{I}^{-1}(\eta_0)B]U_n(\theta_0) \\ &= [Z_n]^T[\mathrm{I}_k - I^{1/2}(\theta_0)B^T\tilde{I}^{-1}(\eta_0)BI^{1/2}(\theta_0)][Z_n] \end{aligned}$$

where $Z_n = I^{-1/2}(\theta_0)U_n(\theta_0) \sim N(0, \mathrm{I}_k)$.

Denote

$$G = \mathrm{I}_k - I^{1/2}(\theta_0)B^T\tilde{I}^{-1}(\eta_0)BI^{1/2}(\theta_0)$$

and note that $\tilde{I}(\eta_0) = BI(\theta_0)B^T$, it is easy to see $G^2 = G$ and thus

$$Rank(G) = tr(G) = k - tr(\tilde{I}^{-1}(\eta_0)BI(\theta_0)B^T) = k - d = s.$$

The proof is completed by the property of chi-square distribution. □

## 11.4 Wald test

- Test on a single parameter $H_0 : \theta = \theta_0$, since the MLE satisfies $\sqrt{n}(\hat{\theta} - \theta) \to N(0, 1/I(\theta))$, the Wald test rejects $H_0$ when

$$W = \frac{(\hat{\theta} - \theta_0)^2}{Var(\hat{\theta})} = n(\hat{\theta} - \theta_0) i(\hat{\theta})(\hat{\theta} - \theta_0) > \chi_1^2(\alpha).$$

- Test(s) on multiple parameters $H_0 : \theta = \theta_0, \theta \in \mathbb{R}^k$. Since $\sqrt{n}(\widehat{\theta} - \theta) \to N_k(0, I^{-1}(\theta))$, it follows that the Wald test that rejects $H_0$ when

$$W_n(\theta_0) = n(\widehat{\theta} - \theta_0)^T i(\widehat{\theta})(\widehat{\theta} - \theta_0) \geq \chi_k^2(\alpha).$$

- For test on $H_0 : g(\theta) = 0_{s \times 1}$ the Wald test rejects $H_0$ when

$$n(g(\hat{\theta}))^T [\nabla g(\hat{\theta}) i^{-1}(\hat{\theta}) (\nabla g(\hat{\theta}))^T]^{-1} (g(\hat{\theta})) > \chi_s^2(\alpha)$$

- $i(\hat{\theta})$ usually takes $-\frac{1}{n} \frac{\partial^2 \ell_n}{\partial \theta \partial \theta^T}|_{\theta = \hat{\theta}}$, $I(\theta_0)$ would also do if easily computed.

Find the Wald test for the hypothesis $H_0 : \theta \in \Theta_0 = \{\theta : \theta_j = \theta_{0,j}, j = 1, \ldots, s\}$.

↑Exam
↓Exam

The $H_0$ can be rewritten as

$$H_0 : g(\theta) = \theta^{(1)} - \theta_0^{(1)} = 0$$

we write the MLE for $\theta \in \Theta$ as $\widehat{\theta}_n = \left(\widehat{\theta}_n^{(1)}, \widehat{\theta}_n^{(2)}\right)$ where $\widehat{\theta}_n^{(1)} = \left(\widehat{\theta}_1, \ldots, \widehat{\theta}_s\right)$ and $\widehat{\theta}_n^{(2)} = \left(\widehat{\theta}_{s+1}, \ldots, \hat{\theta}_k\right)$ and define the Wald statistic as

$$W_n\left(\theta_0^{(1)}\right) = n\left(\widehat{\theta}_n^{(1)} - \theta_0^{(1)}\right)^T \left[I^{11}\left(\widehat{\theta}_n\right)\right]^{-1} \left(\widehat{\theta}_n^{(1)} - \theta_0^{(1)}\right) \xrightarrow{\mathcal{L}} \chi_s^2, \text{under} H_0.$$

where $I^{11}(\theta)$ is the upper diagonal block of $I^{-1}(\theta)$ written as

$$I^{-1}(\theta) = \begin{pmatrix} I^{11}(\theta) & I^{12}(\theta) \\ I^{21}(\theta) & I^{22}(\theta) \end{pmatrix}$$

## 11.5 Score test

- For the simple hypothesis $H_0 : \theta = \theta_0$, Rao's score test or Lagrange Multipliers test is based on the observation that, by the central limit theorem,

$$V_n(\theta_0) \xrightarrow{\mathcal{L}} N\left(0, I(\theta_0)\right),$$

It follows from this that under $H_0$, as $n \to \infty$,

$$R_n\left(\theta_0\right) = V_n^T\left(\theta_0\right) I^{-1}\left(\theta_0\right) V_n\left(\theta_0\right) \xrightarrow{\mathcal{L}} \chi_k^2.$$

The test rejects $H_0$ when $R_n\left(\theta_0\right) \geq \chi_k^2(\alpha)$.

- Equivalently, letting $H$ be the Lagrangian:

$$H = \ell_n(\theta) - \lambda'\left(\theta - \theta^0\right),$$

the first-order conditions are:

$$\frac{\partial \ell_n}{\partial \theta} = \lambda; \quad \theta = \theta^0,$$

so $\lambda/\sqrt{n} = V_n(\theta_0) \xrightarrow{\mathcal{L}} N\left(0, I(\theta_0)\right)$. Thus, $R_n = \lambda^T I^{-1}(\theta_0)\lambda/n$.

- For test on $H_0 : g(\theta) = 0_{s\times 1}$, the statistic employed in the score test is based on the ML estimate $\widehat{\theta}_n^R$ that is obtained from the solution of the constrained optimization problem

$$\widehat{\theta}_n^R = \arg \max_{\theta \in H_0} \ell_n(\theta)$$

The test statistic, called score statistic, is

$$LM_n = U_n(\hat{\theta}_n^R)^\top I^{-1}(\widehat{\theta}_n^R) U_n(\widehat{\theta}_n^R) \xrightarrow{\mathcal{L}} \chi^2_s, \ under \ H_0$$

- Note that $I^{-1}(\theta_0)$ depends on the unknown parameter value $\theta_0$. We can evaluate the terms involved in $\theta_0$ at the constrained MLE, $\widehat{\theta}_n^R$ to get a usable statistic. We can approximate $I(\theta_0)$ with either $-\frac{1}{n}\nabla^2 \ell_n(\widehat{\theta}_n^R)$ or $\frac{1}{n}\nabla \ell_n(\widehat{\theta}_n^R)\nabla^\top \ell_n(\widehat{\theta}_n^R)$.

*Proof.* FOC's are given by

$$\nabla \ell_n(\widehat{\theta}_n^R) + \nabla g(\widehat{\theta}_n^R)\lambda = 0$$
$$g(\widehat{\theta}_n^R) = 0$$

First order Taylor expansions of $g(\widehat{\theta}_n)$ and $g(\widehat{\theta}_n^R)$ around $\theta_0$ gives, ignoring $o_p(1)$ terms,

$$\sqrt{n}g(\widehat{\theta}_n) = \sqrt{n}g(\theta_0) + \nabla g'(\theta_0)\sqrt{n}(\widehat{\theta}_n - \theta_0)$$
$$\sqrt{n}g(\widehat{\theta}_n^R) = \sqrt{n}g(\theta_0) + \nabla g'(\theta_0)\sqrt{n}(\widehat{\theta}_n^R - \theta_0)$$

Note that $g(\widehat{\theta}_n^R) = 0$, we have

$$\sqrt{n}g(\widehat{\theta}_n) = \nabla g'(\theta_0)\sqrt{n}(\widehat{\theta}_n - \widehat{\theta}_n^R)$$

On the other hand, taking first order Taylor series expansions of $\nabla \ell_n(\widehat{\theta}_n)$ and $\nabla \ell_n(\widehat{\theta}_n^R)$ around $\theta_0$ gives, ignoring $o_p(1)$ terms,

$$0 = \nabla \ell_n(\widehat{\theta}_n) = \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\theta_0)(\widehat{\theta}_n - \theta_0) \Rightarrow$$

$$0 = U_n(\widehat{\theta}_n) = U_n(\theta_0) - I(\theta_0)\sqrt{n}(\widehat{\theta}_n - \theta_0),$$

note that $-\nabla^2 \ell_n(\theta_0) \to I(\theta_0)$ by the law of large numbers. Similarly,

$$U_n(\widehat{\theta}_n^R) = \nabla \ell_n(\widehat{\theta}_n^R) = U_n(\theta_0) - I(\theta_0)\sqrt{n}(\widehat{\theta}_n^R - \theta_0)).$$

Taking the difference, we have

$$U_n(\widehat{\theta}_n^R) = I(\theta_0)\sqrt{n}(\widehat{\theta}_n - \widehat{\theta}_n^R).$$

Hence,

$$\sqrt{n}(\widehat{\theta}_n - \widehat{\theta}_n^R) = I^{-1}(\theta_0)U_n(\widehat{\theta}_n^R).$$

Therefore,

$$\sqrt{n}g(\widehat{\theta}_n) = \nabla g'(\theta_0)I^{-1}(\theta_0)U_n(\widehat{\theta}_n^R).$$

We then deduce

$$\begin{aligned}\sqrt{n}g(\widehat{\theta}_n) &= -\nabla g'(\theta_0)I^{-1}(\theta_0)\nabla g(\widehat{\theta}_n^R)\frac{\lambda}{\sqrt{n}} \\ &\to -\nabla g'(\theta_0)I^{-1}(\theta_0)\nabla g(\theta_0)\frac{\lambda}{\sqrt{n}},\end{aligned}$$

since $\widehat{\theta}_n^R \to \theta_0$ hence $g(\widehat{\theta}_n^R) \to g(\theta_0)$ under the null hypothesis. By $\sqrt{n} g(\widehat{\theta}_n) \xrightarrow{\mathcal{L}} N(0, \nabla g'(\theta_0) I^{-1}(\theta_0) \nabla g(\theta_0))$, we have

$$\frac{\lambda}{\sqrt{n}} \xrightarrow{\mathcal{L}} N\left(0, \left(\nabla g'(\theta_0) I^{-1}(\theta_0) \nabla g(\theta_0)\right)^{-1}\right)$$

Again, forming the quadratic form of the normal random variables, we obtain

$$\frac{1}{n} \lambda^\top \left(\nabla g'(\theta_0) I^{-1}(\theta_0) \nabla g(\theta_0)\right) \lambda \xrightarrow{\mathcal{L}} \chi_s^2 \quad \text{under } H_0$$

Alternatively, another form of the test statistic is given by

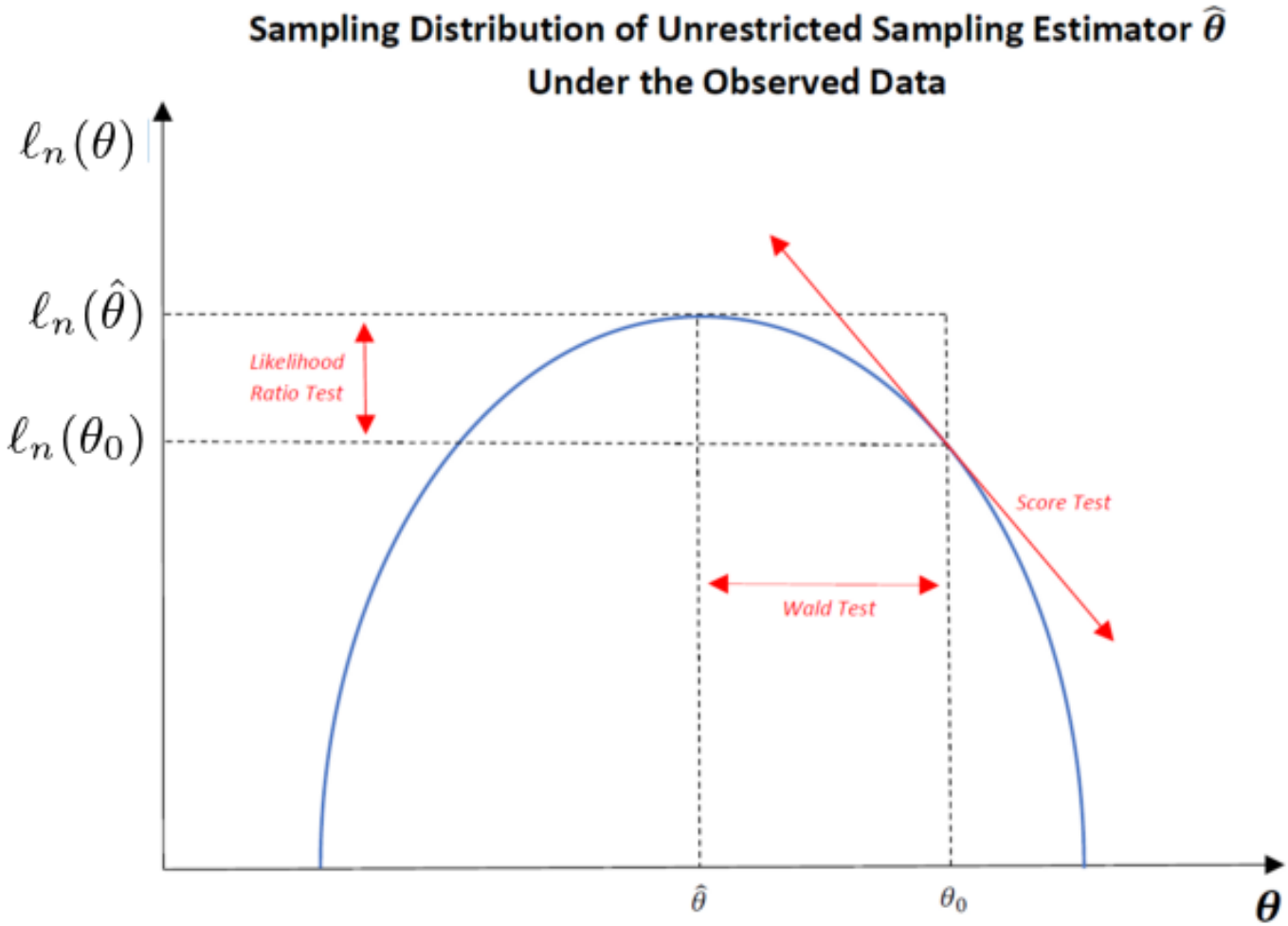
$$U_n^\top(\widehat{\theta}_n^R) I^{-1}(\theta_0) U_n(\widehat{\theta}_n^R) \xrightarrow{\mathcal{L}} \chi_s^2 \quad \text{under } H_0$$

□

## Summary

All of the test statistics mentioned here are asymptotically equivalent to the LRT (Wald and Rao score alike).

- an LRT requires computation of an MLE, $\widehat{\theta}_n$, and a restricted MLE, $\widehat{\theta}_n^R$, and it is a function of the distance between $\ell_n(\hat{\theta}_n)$ and $\ell_n(\widehat{\theta}_n^R)$
- a Wald test requires only computation of the MLE $\widehat{\theta}_n$, and it is a function of the distance between $g(\hat{\theta}_n)$ and $g(\theta_0) = 0$.
- a Rao score test requires only computation of the restricted MLE $\widehat{\theta}_n^R$ and measures how far the Lagrangian multiplier is from zero.
- As shown above, Wald, LM and LR test are asymptotically equivalent with $\chi_s^2$.
- LRT is usually more powerful than Wald and Rao score tests under small or moderate sample size.




LRT, Wald and Score tests for $H_0 : \theta = \theta_0$