Lecture 8: Decision theory and likelihood ratio test

8.1 The Decision Theoretic Framework

In practical applications, mathematical frameworks often guide decision-making by helping choose among alternatives. It is generally considered rational to select options with the most favorable outcomes under uncertainty; e.g., those that maximize expected profits, minimize expected losses, maximize expected sales, or minimize expected costs. We formulate the decision process when we are given data.

(a) Data, sample space and model: by observing a sample X, we model the data generating scheme $X \sim P_\theta$ with $\theta \in \Theta$. E.g., $X \sim \mathcal{N}(\mu, \sigma^2)$.

(b) Action space $\mathcal{A}$: a collection of possible actions. E.g., $\mathcal{A} = \{\text{reject the null }, \text{not reject the null}\}$ in the hypothesis testing.

(c) Decision rule δ: let $\mathcal{X}$ is a set of all possible values of data X, a decision rule is a function $\mathcal{X} \to \mathcal{A}$ that outputs an action $a = \delta(X)$ given data X.

(d) Criterion: loss function ℓ and risk function R.

Example 8.1 Suppose a company is examining land for oil drilling. The company identifies land with the possibility of containing a rock formation, which they believe is caused by an oil reserve. Based on if this formation exists, they make an educated guess about whether the land contains oil. Suppose that we observe $X \sim P_\theta$, where $\theta \in \Theta = \{\theta_1^{\text{oil}}, \theta_2^{\text{no oil}}\}$ and $\mathcal{X} = \{0, 1\}$ (i.e., $X = 1$ means rock formation)

	$P_\theta(X = 0)$	$P_\theta(X = 1)$	total
θ_1^{oil}	0.3	0.7	1
$\theta_2^{\text{no oil}}$	0.6	0.4	1

After observing $X = x$, the company needs to take an action in $\mathcal{A} = \{a_1 = \text{drill}, a_2 = \text{sell}, a_3 = \text{sell partial right}\}$.

(i) Decision rule: we have 9 different decision rules δ_i for $i = 1, \ldots, 9$ as the table below.

i	1	2	3	4	5	6	7	8	9
x = 0	a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3
x = 1	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3

Here, δ_1 represents "Take action a_1 regardless of the value of X," δ_2 corresponds to "Take action a_1, if $X = 0$; take action a_2, if $X = 1$," and so on.

(ii) Loss function, risk function, and risk points: suppose the company internally evaluates the loss function $l(\theta, a)$ as the table below.

	a_1 = drill	a_2 = sell	a_3 = sell partial rights
θ_1^{oil}	0	10	5
$\theta_2^{\text{no oil}}$	12	1	6

The risk of a decision function δ at θ is $R(\theta, \delta) = \mathbb{E}[l(\theta, \delta(X))]$. There are two ways to view this expectation:

$$\begin{aligned}
R(\theta, \delta) &= \mathbb{E}_\theta[l(\theta, \delta(X))] \\
&= \sum_{x=0}^{1} l(\theta, \delta(x)) \cdot P_\theta(X = x) && \text{(option 1: } X \text{ is R.V.)} \\
&= l(\theta, \delta(0)) \cdot P_\theta(X = 0) + l(\theta, \delta(1)) \cdot P_\theta(X = 1) \\
&= \sum_{\delta(x) \in \mathcal{A}} l(\theta, \delta(x)) \cdot P_\theta(\delta(X) = \delta(x)) && \text{(option 2: } \delta(X) \text{ is R.V.)} \\
&= l(\theta, a_1)P[\delta(X) = a_1] + l(\theta, a_2)P[\delta(X) = a_2] + l(\theta, a_3)P[\delta(X) = a_3]
\end{aligned}$$

For instance, if the true parameter is θ_1 (oil), the risk of the decision δ_2 is

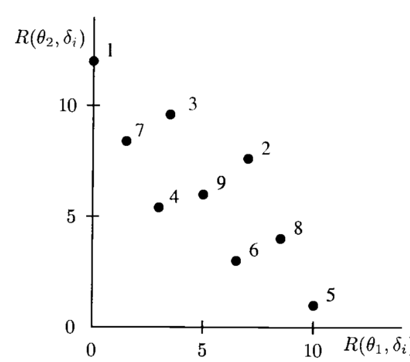
$$R(\theta_1, \delta_2) = 0(0.3) + 10(0.7) = 7.$$

Likewise if the true parameter is θ_2 (no oil), the risk of the decision δ_2 is

$$R(\theta_2, \delta_2) = 12(0.6) + 1(0.4) = 7.6.$$

(iii) Depending on the true parameter, we have different risk value. Since there exists two possible true parameter values, we can plot the **risk point** of a decision δ as a tuple $(R(\theta_1, \delta), R(\theta_2, \delta))$. The table calculates the values of the 9 risk points corresponding to each decision rule.

i	1	2	3	4	5	6	7	8	9
$R(\theta_1, \delta_i)$	0	7	3.5	3	10	6.5	1.5	8.5	5
$R(\theta_2, \delta_i)$	12	7.6	9.6	5.4	1	3	8.4	4.0	6



The figure above shows the 9 risk points corresponding each decision rule. Which do you think is the best decision rule?

8.2 Statistical hypothesis testing

Remark 8.1 Let's think about the process of modeling an observed phenomenon. Suppose you need to assess whether the current hypothesis is met. In such cases, we need to make decisions whether to accept the hypothesis or to reject the hypothesis (thus take the alternative hypothesis) based on the data and evidence. This process is called **statistical hypothesis testing**.

Example 8.2 In a pharmaceutical company, the time until a new pill induces a sedative effect is known to follow a normal distribution with a mean time of 30 minutes and a standard deviation of 5 minutes. A researcher wishes to determine whether a newly developed pill has a faster sedative effect than the existing pill. The researcher performs a study by administering the new pill to 100 participants and measuring the time until the sedative effect occurs. Let the sample mean time be $\bar{X}$. If the researcher believes that the new pill is faster than the existing one, how should they assess whether this claim is statistically significant?

Solution. Let the null hypothesis H_0 represent the claim that the new pill has the same sedative effect time as the existing pill. This can be stated as

$$H_0 : \mu = 30$$

The alternative hypothesis H_1, which represents the claim that the new pill is faster, is

$$H_1 : \mu < 30$$

In this case, we are interested in whether the new pill leads to a faster sedative effect. The null hypothesis H_0 assumes that the new pill does not reduce the sedative time, while the alternative hypothesis H_1 suggests that the new pill reduces the sedative time. To test these hypotheses, the researcher uses the sample mean $\bar{X}$. Suppose the observed mean is $\bar{X} = \bar{x} = 29$. From our test statistic, $\bar{X}$, we can transform it into Z, where under the null hypothesis,

$$Z = \frac{\bar{X} - 30}{\frac{5}{\sqrt{100}}} = -2$$

is known to follow $Z \sim N(0, 1)$. Thus, the probability of obtaining this result or something more extreme [in the direction of the alternative hypothesis] under the null hypothesis (i.e., **significance probability**) is:

$$P(\bar{X} \leq 29 | H_0) = P\left(Z \leq \frac{29 - 30}{\frac{5}{\sqrt{100}}}\right) = P(Z \leq -2) = 0.023$$

Thus, if the sample mean $\bar{X} = 29$ is observed, the probability of obtaining this sedative time or faster under the null hypothesis is very small (about 2.3%). This interpretation is our *decision rule* based on this significance probability (which is called the "p-value," p). Another way to interpret the p-value is, what is the probability that your observed data is due to random chance alone (i.e., under H_0)? If this probability is small, then perhaps this is evidence that your observations are *not* due to random chance, but due to some other reason (i.e., H_1). For our example, $p = 0.023$ suggests that the null hypothesis $H_0 : \mu = 30$ may not be true (i.e., the model with $\mu = 30$ "disagrees" with the observations), and we might reject it in favor of the alternative hypothesis $H_1 : \mu < 30$.

In this case, the null hypothesis represents the claim that there is no difference in the sedative effect time, while the alternative hypothesis suggests that the new pill is faster. More generally, if the value of the test statistic here ($\bar{X}$) is ≤ 29, we say it is in the **rejection region**. ■

Remark 8.2 Summary: hypothesis testing relies on test statistics to make decisions about whether to reject or accept a hypothesis based on the evidence provided by the data.

Definition 8.1 In general, when one wishes to refute a certain hypothesis based on the observed results, the initial hypothesis is called the **null hypothesis** (H_0), and the hypothesis proposed as an alternative is called the **alternative hypothesis** (H_1). Testing refers to using a sample to decide whether to reject or fail to reject the null hypothesis. Thus, **testing is expressed in terms of the rejection region of the null hypothesis**, and a statistical quantity used to indicate whether the null hypothesis is rejected or not, such as the sample mean, is called the **test statistic**.

Definition 8.2 In hypothesis testing, two types of errors can occur when either rejecting or failing to reject the null hypothesis:

Test Result	Null Hypothesis is True (H_0)	Alternative Hypothesis is True (H_1)
Accept H_0	*correct decision*	Type II Error
Reject H_0	Type I Error	*correct decision*

Remark 8.3 It is desirable to minimize the probability of making such errors as much as possible. Usually we put constraints on Type I error (i.e., the error to reject the null hypothesis when the null hypothesis is true). Tests are designed to minimize the probability of committing a Type I error with a predetermined **significance level**: the **maximum allowable probability of committing a Type I error** (α). Commonly used significance levels are $\alpha = 0.10$, $\alpha = 0.05$, and $\alpha = 0.01$.

Definition 8.3 The **significance level** refers to the maximum allowable probability of rejecting the null hypothesis H_0 when it is true, that is, committing a Type I error. In other words, it is:

$$P(\text{Rejecting } H_0 \mid H_0 \text{ is true}) \leq \alpha$$

A test conducted under significance level α is called a *test of significance level α.*

Example 8.3 In Example 8.2, the null and alternative hypotheses are

$$H_0 : \mu = 30 \quad \text{vs.} \quad H_1 : \mu < 30$$

and

$$\bar{X} \sim N\left(\mu,\ \frac{5^2}{100}\right), \quad Z = \frac{\bar{X} - \mu}{5/\sqrt{100}} \sim N(0, 1).$$

For these tests, we define a significance level, generally assumed to be, $\alpha = 0.05$ which has the following interpretation for this example

$$P\left(\frac{\bar{X} - 30}{5/\sqrt{100}} \leq z_{0.05} \mid H_0\right) = 0.05 = \alpha.$$

Thus, the test for "$\bar{X} \leq 29$" becomes:

$$P(\bar{X} \leq 29 \mid H_0) = P\left(\frac{\bar{X} - 30}{5/\sqrt{100}} \leq -2\right) \approx 0.023 \leq 0.05.$$

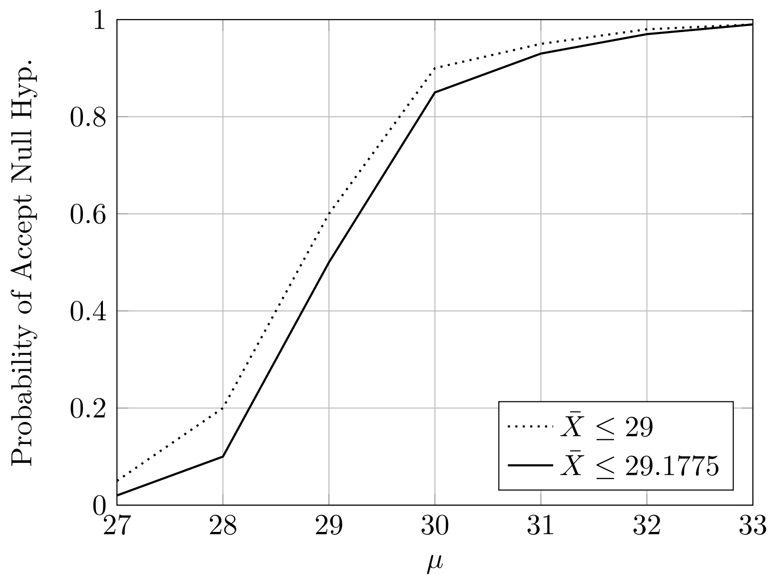
We can determine the rejection region for a test of significance at level $\alpha = 0.05$ to be

$$\begin{aligned}
\frac{\bar{X} - 30}{5/\sqrt{100}} &\leq z_{0.05} \\
\bar{X} &\leq 30 + z_{0.05} \times \frac{5}{\sqrt{100}} \\
\bar{X} &\leq 30 - 1.645 \times \frac{5}{\sqrt{100}} \\
\bar{X} &\leq 29.1775.
\end{aligned}$$

Example 8.4 Continuing to build off of Example 8.2, let us shift focus to determining the Type II error. In this case, we are interested in the probability of accepting the null hypothesis given the truth is H_1 (i.e., $\mu < 30$). Consider two possible rejection regions: (a) $\bar{X} \leq 29$, and (b) $\bar{X} \leq 29.1775$

$$\begin{aligned}
&(a) \quad P_\mu(\bar{X} \leq 29) = P\left(\frac{\bar{X} - \mu}{5/\sqrt{100}} \leq \frac{29 - \mu}{5/\sqrt{100}}\right) = \Phi(58 - 2\mu), \\
&(b) \quad P_\mu(\bar{X} \leq 29.1775) = P\left(\frac{\bar{X} - \mu}{5/\sqrt{100}} \leq \frac{29.1775 - \mu}{5/\sqrt{100}}\right) = \Phi(58.355 - 2\mu)
\end{aligned}$$

where Φ is the CDF of a standard normal distribution. By the figure below we see that the test with rejection region $\bar{X} \leq 29$ has higher the Type II error probability than the test with rejection region $\bar{X} \leq 29.1775$. Having a rejection region as large as possible while still maintaining a significance level α reduces the probability of committing a Type II error. In other words, for a significance level α, we select the test with $P(\text{Reject } H_0 \mid H_0) = \alpha$, not $P(\text{Reject } H_0 \mid H_0) < \alpha$.



Definition 8.4 For a given hypothesis test

$$H_0 \ : \ \theta \in \omega_0 \quad \text{v.s.} \quad H_1 \ : \ \theta \in \omega_1$$

we say a critical/rejection region, C, is of **size $\boldsymbol{\alpha}$** if

$$\alpha = \max_{\theta \in \omega_0} P_\theta[(X_1, \ldots, X_n) \in C].$$

Subsequently, the **power function** of a critical/rejection region, C, is defined as

$$\gamma_C(\theta) = P_\theta[(X_1, \ldots, X_n) \in C], \quad \theta \in \omega_1.$$

Remark 8.4 If we represent the probability of a Type II error as $P(\text{accept } H_0 \mid H_1 \text{ true})$, then we want to minimize this probability. This is the same as maximizing its compliment, namely, maximizing $1 - P(\text{accept } H_0 \mid H_1 \text{ true}) = P(\text{reject } H_0 \mid H_1 \text{ true})$ which means maximizing the **power**. That is to say, maximizing the "power" to identify the alternative hypothesis.

Remark 8.5 If two critical regions C_1 and C_2 are both size α, then C_1 is "better" than C_2 if $\gamma_{C_1}(\theta) \geq \gamma_{C_2}(\theta)$, for all $\theta \in \omega_1$.

8.3 Likelihood ratio test

Remark 8.6 **(Background/Motivation)** Consider the following hypotheses

$$H_0 : \theta \in \Omega_0 \quad \text{vs.} \quad H_1 : \theta \in \Omega_1 \quad (\Omega_0 \cap \Omega_1 = \emptyset, \Omega_0 \cup \Omega_1 = \Omega)$$

which we aim to test based on an observed sample $X_1, \ldots, X_n$ assumed to be i.i.d. from a pdf $f(x; \theta)$, where $\theta \in \Omega$ is a parameter. For each hypothesis, the observed results $(x_1, \ldots, x_n)$ are compared based on their likelihood (because recall likelihoods are functions of *parameters*). We can conduct a hypothesis test by performing the **likelihood ratio test**. Recall, the likelihood function is given by

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i; \theta), \quad \theta \in \Omega,$$

and we can study likelihood ratio is defined as

$$\frac{\max_{\theta \in \Omega_0} L(\theta; \boldsymbol{x})}{\max_{\theta \in \Omega_1} L(\theta; \boldsymbol{x})}.$$

If this ratio is small, it suggests that the alternative hypothesis H_1 is more likely than the null hypothesis H_0 because the likelihood of $\theta \in \Omega_1$ is greater, given the data. This would lead us to reject H_0. However, doing this comparison between $\theta \in \Omega_0$ versus $\theta \in \Omega_1$ does not answer the question of "Do we reject H_0, or fail to reject H_0?" In other words, the focus of the question before deals solely with the feasibility of H_0, thus we should compare $\theta \in \Omega_0$ to $\theta \in \Omega$ (i.e., compare H_0 across the entire parameter space). Also, note that doing the maximization over $\theta \in \Omega_1$ is often less intuitive than doing the maximum over $\theta \in \Omega_0$ and $\theta \in \Omega$.

Lemma 8.1 *Suppose $X_1, \ldots, X_n \overset{iid}{\sim} f_X(x \mid \theta)$. Assume that $\theta = \theta_0$ is the true parameter and that $\mathbb{E}_{\theta_0}[f_X(x_i \mid \theta)/f_X(x_i \mid \theta_0)]$ exists. Under regularity assumptions R0 and R1 (from Lecture 5),*

$$\lim_{n \to \infty} P_{\theta_0}[L(\theta_0 \mid \boldsymbol{X}) > L(\theta \mid \boldsymbol{X})] = 1, \quad \forall \theta \neq \theta_0$$

where $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Definition 8.5 Let $X_1, \ldots, X_n \overset{iid}{\sim} f_X(x \mid \theta)$ for $\theta \in \Omega$. Consider the following hypothesis

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

Let $\hat{\theta}$ denote the maximum likelihood estimate of $\theta \in \Omega$. The **likelihood ratio** is defined as

$$\Lambda = \frac{L(\theta_0 \mid X_1, \ldots, X_n)}{L(\hat{\theta} \mid X_1, \ldots, X_n)} \quad \text{or simply} \quad \Lambda = \frac{L(\theta_0)}{L(\hat{\theta})}$$

By Lemma 8.1, $\Lambda \leq 1$. Therefore, the **likelihood ratio test (LRT)** says for a specified significance level α, the decision rule is

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } \Lambda \le c$$

where c is chosen such that $\alpha = P_{\theta_0}(\Lambda \le c)$, $c \in [0, 1]$.

Remark 8.7 Much like when finding MLEs, it is often easier to work with log-likelihoods as opposed to likelihoods. Therefore, we can alternatively write the likelihood ratio as

$$\lambda = l(\theta_0) - l(\hat{\theta})$$

However, it is worth noting that more often it is written as

$$\lambda^* = -2[l(\theta_0) - l(\hat{\theta})]$$

because this -2 factor ensures "nice" asymptotic properties. The rejection region in either case is then given by

$$\lambda \le c \quad \text{and} \quad \lambda^* \ge c, \quad \textbf{respectively}$$

Definition 8.6 **(More general LRT)** Let $X_1, \ldots, X_n \overset{iid}{\sim} f_X(x \mid \theta)$ for $\theta \in \Omega$. At the significance level α, to likelihood ratio test for

$$H_0 : \theta \in \Omega_0 \quad \text{vs.} \quad H_1 : \theta \in \Omega_1 \quad (\Omega_0 \cap \Omega_1 = \emptyset, \Omega_0 \cup \Omega_1 = \Omega).$$

is defined as follows.

(a) The rejection region is defined as

$$C_\alpha = \left\{ \boldsymbol{x} : \frac{\max_{\theta \in \Omega_0} L(\theta; \boldsymbol{x})}{\max_{\theta \in \Omega} L(\theta; \boldsymbol{x})} \le c, \ c \in [0, 1] \right\} \quad \text{or} \quad C_\alpha = \left\{ \boldsymbol{x} : -2\left(l(\hat{\theta}_0; \boldsymbol{x}) - l(\hat{\theta}; \boldsymbol{x}) \right) \ge c, \ c \in [0, \infty) \right\}.$$

where $\lambda^* = -2\left(l(\hat{\theta}_0; \boldsymbol{x}) - l(\hat{\theta}; \boldsymbol{x}) \right)$.

(b) The threshold c is determined such that

$$\max_{\theta \in \Omega_0} P_\theta \left((X_1, \ldots, X_n)^T \in C_\alpha \right) = \alpha.$$

Example 8.5 Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ where $\theta \in (-\infty, \infty)$ and $\sigma^2 > 0$. Suppose σ^2 is known, and we want to test

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

Construct a likelihood ratio test for the aforementioned hypotheses.

(a) Find Λ (or λ) (recall $\hat{\theta}_{MLE} = \bar{X}$).
Solution. First, recall the pdf of $N(\theta, \sigma^2)$ as

$$f(x \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

for $x \in (-\infty, \infty)$. The log-likelihood is thus given as

$$l(\theta \mid X_1, \ldots, X_n, \sigma^2) = \sum_{i=1}^{n} \frac{-1}{2} \log\left(2\pi\sigma^2\right) - \frac{(x_i - \theta)^2}{2\sigma^2} = \frac{-n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2$$

Now, we can compute the likelihood ratio as

$$\begin{aligned}
\lambda &= l(\theta_0) - l(\hat{\theta}) \\
&= \frac{-n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta_0)^2 + \frac{n}{2} \log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 - (x_i - \theta_0)^2
\end{aligned}$$

Recall, we want to simplify this until it is in the form of something that we know the distribution of (in order to do part (b)).

$$\begin{aligned}
\lambda &= \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 - (x_i - \theta_0)^2 \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \bar{x}^2 - 2x_i\bar{x} + 2x_i\theta_0 - \theta_0^2 \\
&= \frac{1}{2\sigma^2} \left\{ n\bar{x}^2 - 2\bar{x} \sum_{i=1}^{n} x_i + 2\theta_0 \sum_{i=1}^{n} x_i - n\theta_0^2 \right\} \\
&= \frac{1}{2\sigma^2} \left\{ n\bar{x}^2 - 2n\bar{x}^2 + 2n\theta_0\bar{x} - n\theta_0^2 \right\} \\
&= \frac{1}{2\sigma^2} \left\{ -n\bar{x}^2 + 2n\theta_0\bar{x} - n\theta_0^2 \right\} \\
&= -\frac{n}{2\sigma^2} \left\{ \bar{x}^2 - 2\theta_0\bar{x} + \theta_0^2 \right\} \\
&= -\frac{n}{2\sigma^2} (\bar{x} - \theta_0)^2 \\
&= \frac{-(\bar{x} - \theta_0)^2}{2\sigma^2/n}
\end{aligned}$$

Using Remark 8.7,

$$\lambda^* = -2[l(\theta_0) - l(\hat{\theta})] = \frac{(\bar{x} - \theta_0)^2}{\sigma^2/n} = \left[\frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}}\right]^2$$

Writing this as an estimator (which is necessary for computing the region region in part (b))

$$\lambda^* = \left[\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}\right]^2 \overset{H_0}{\sim} \chi^2(1)$$

because $\frac{\bar{X}-\theta_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$. ■

(b) Find the value of c such that $\alpha = P_{\theta_0}(\Lambda \leq c)$, or $\alpha = P_{\theta_0}(\lambda \leq c)$.

Solution. The value of c can be found by as follows:

$$\begin{aligned}
\alpha &= P_{\theta_0}(\lambda^* \geq c) \\
&= P_{\theta_0}\left(\left[\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}\right]^2 \geq c\right) \\
&= 1 - P_{\theta_0}\left(\left[\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}\right]^2 < c\right) \\
&= 1 - F_Y(c)
\end{aligned}$$

where $Y = \left[\frac{\bar{X}-\theta_0}{\sigma/\sqrt{n}}\right]^2 \sim \chi^2(1)$. Thus, $c = F_Y^{-1}(1-\alpha)$ (i.e., the $1-\alpha$ quantile of a $\chi^2(1)$). ■