Chapter 8

Hypothesis Testing

8.1 Introduction

Definition 8.1.1 *A hypothesis is a statement about a population parameter.*

The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true.

Definition 8.1.2 *The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by H_0 and H_1, respectively.*

Definition 8.1.3 *A hypothesis testing procedure or hypothesis test is a rule that specifies:*

i. For which sample values the decision is made to accept H_0 as true.

ii. For which sample values H_0 is rejected and H_1 is accepted as true.

The subset of the sample space for which H_0 will be rejected is called the rejection region or critical region. The complement of the rejection region is called the acceptance region.

8.2 Methods of Finding Tests

8.2.1 Likelihood Ratio Tests

Definition 8.2.1 *The likelihood ratio test statistic for testing* $H_0 : \theta \in \Theta_0$ *versus* $H_1 : \theta \in \Theta_0^c$ *is*

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\Theta} L(\theta|\boldsymbol{x})}.$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$*, where* c *is any number satisfying* $0 \leq c \leq 1$*.*

The rationale behind LRTs can be easily understood in a situation in which $f(x|\theta)$ is the pmf of a discrete random variable. In this case, the numerator of $\lambda(\boldsymbol{x})$ is the maximum probability of the observed sample, the maximum being computed over parameters in the null hypothesis. The denominator of $\lambda(\boldsymbol{x})$ is the maximum probability of the observed sample over all possible parameters. The ratio of these two maxima is small if there are parameter points in the alternative hypothesis for which the observed sample is much more likely than for any parameter point in the null hypothesis. In this situation, the LRT criterion says H_0 should be rejected and H_1 accepted as true. Suppose $\hat{\theta}$, an MLE of θ, exists; $\hat{\theta}$ is obtained by doing an unrestricted maximization of $L(\theta|\boldsymbol{x})$. We can also consider the MLE of θ, call it $\hat{\theta}_0$, obtained by doing a restricted maximization, assuming Θ_0 is the parameter space. Then, the LRT statistic is

$$\lambda(\boldsymbol{x}) = \frac{L(\hat{\theta}_0|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})}.$$

Example 8.2.1 *(***Normal LRT***) Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, 1)$ population. Considering testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Since there is only one value of θ specified by H_0, the numerator of $\lambda(\boldsymbol{x})$ is $L(\theta_0|\boldsymbol{x})$. Recall that the MLE of θ is $\bar{X}$. Thus the denominator of $\lambda(\boldsymbol{x})$ is $L(\bar{x}|\boldsymbol{x})$. So the LRT statistic is*

$$\begin{aligned}\lambda(\boldsymbol{x}) &= \frac{(2\pi)^{-n/2}\exp[-\sum_{i=1}^{n}(x_i-\theta_0)^2/2]}{(2\pi)^{-n/2}\exp[-\sum_{i=1}^{n}(x_i-\bar{x})^2/2]} \\ &= \exp[-n(\bar{x}-\theta_0)^2/2].\end{aligned}$$

The rejection region, $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$, can be written as

$$\{\boldsymbol{x} : |\bar{x}-\theta_0| \geq \sqrt{-2(\log c)/n}.$$

Thus, the LRTs are just those tests that reject $H_0 : \theta = \theta_0$ if the sample mean differs from the hypothesized value θ_0 by more than a specified amount.

Given the intuitive notion that all the information about θ in $\boldsymbol{x}$ is contained in $T(\boldsymbol{x})$, a sufficient statistic for θ, the test based on T should be as good as the test based on the complete sample $\boldsymbol{X}$. In fact the tests are equivalent.

Theorem 8.2.1 *If $T(\boldsymbol{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(\boldsymbol{x})$ are the LRT statistics based on T and $\boldsymbol{X}$, respectively, then $\lambda^*(T(\boldsymbol{x})) = \lambda(\boldsymbol{x})$ for every $\boldsymbol{x}$ in the sample space.*

PROOF: From the Factorization Theorem, the pdf or pmf of $\boldsymbol{X}$ can be written as $f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$, where $g(t|\theta)$ is the pdf or pmf of T and $h(\boldsymbol{x})$ does not depend on θ. Thus

$$\begin{aligned}\lambda(\boldsymbol{x}) &= \frac{\sup_{\Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\Theta} L(\theta|\boldsymbol{x})} = \frac{\sup_{\Theta_0} g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})}{\sup_{\Theta} g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})} \\ &= \frac{\sup_{\Theta_0} g(T(\boldsymbol{x})|\theta)}{\sup_{\Theta} g(T(\boldsymbol{x})|\theta)} = \frac{\sup_{\Theta_0} L^*(\theta|T(\boldsymbol{x}))}{\sup_{\Theta} L(\theta|T(\boldsymbol{x}))} = \lambda^*(T(\boldsymbol{x})).\end{aligned}$$

The theorem is proved. □

Example 8.2.2 *(***LRT and sufficiency***) In Example 8.2.1, we can recognize that $\bar{X}$ is a sufficient statistic for θ, and $\bar{X} \sim N(\theta, \frac{1}{n})$. Based on that it is easy to conclude that a likelihood ratio test of $H_0 : \theta = \theta_0$ versus $H_0 : \theta \neq \theta_0$ rejects H_0 for large values of $|\bar{X} - \theta_0|$.*

8.2.2 Bayesian Tests

Recall that $\pi(\theta|\boldsymbol{x})$ is the posterior distribution of θ. The posterior probability $\pi(\theta \in \Theta_0|\boldsymbol{x}) = P(H_0 \text{ is true}|\boldsymbol{x})$ and $\pi(\theta \in \Theta_0^c|\boldsymbol{x}) = P(H_1 \text{ is true}|\boldsymbol{x})$ can be computed for hypothesis testing. H_0 will be accepted if $\pi(\theta \in \Theta_0^c|\boldsymbol{X}) \geq P(H_1 \text{ is true}|\boldsymbol{X})$ and will be rejected otherwise. In the terminology of the previous sections, the test statistic, a function of the sample, is $P(\theta \in \Theta_0^c|\boldsymbol{X})$ and the rejection region is $\{\boldsymbol{x} : P(\theta \in \Theta_0^c|\boldsymbol{x}) > 1/2\}$. Alternatively, if the Bayesian hypothesis tester wishes to guard against falsely rejecting H_0, he may decide to reject H_0 only if $P(\theta \in \Theta_0^c|\boldsymbol{X})$ is greater than some large number, 0.99 for example.

Example 8.2.3 *(***Normal Bayesian test***) Let $X_1, \ldots, X_n$ be a random sample from $N(\theta, \sigma^2)$ and let the prior distribution on θ be $N(\mu, \tau^2)$, where σ^2, μ and τ^2 are known. Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. The posterior $\pi(\theta|\boldsymbol{x})$ is normal with mean $(n\tau^2\bar{x} + \sigma^2\mu)/(n\tau^2 + \sigma^2)$ and variance $\sigma^2\tau^2/(n\tau^2 + \sigma^2)$.*

If we decide to accept H_0 if and only if $P(\theta \in \Theta_0|\boldsymbol{X}) \geq P(\theta \in \Theta_0^x|\boldsymbol{X})$, then we will accept H_0 if and only if

$$\frac{1}{2} \leq P(\theta \in \Theta_0|\boldsymbol{X}) = P(\theta \leq \theta_0|\boldsymbol{X}).$$

Therefore h_0 will be accepted as true if

$$\bar{X} \leq \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2}$$

and H_1 will be accepted as true otherwise.

8.2.3 Union-Intersection and Intersection-Union tests

The union-intersection method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection, say

$$H_0 : \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma.$$

Here Γ is an arbitrary index set that may be finite or infinite, depending on the problem. Suppose that tests are available for each of the problems of testing $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$. Say the rejection region for the test of $H_{0\gamma}$ is $\{\boldsymbol{x} : T_\gamma(\boldsymbol{x}) \in R_\gamma\}$. Then the rejection region for the union-intersection test is

$$\bigcup_{\gamma \in \Gamma} \{\boldsymbol{x} : T_\gamma(\boldsymbol{x}) \in R_\gamma\}.$$

The rationale is simple. If any one of the hypotheses $H_{0\gamma}$ is rejected, then H_0 must also be rejected.

Example 8.2.4 *(***Normal union-intersection test***) Let $X_1, \ldots, X_n$ be a random sample from $N(\theta, \sigma^2)$. Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, where μ_0 is a specified number. We can write H_0 as the intersection of two sets,*

$$H_0 : \{\mu : \mu \leq \mu_0\} \cap \{\mu : \mu \geq \mu_0\}.$$

The LRT of $H_{0L} : \mu \leq \mu_0$ versus $H_{1L} : \mu > \mu_0$ is: reject H_{0L} if $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq t_L$. The LRT of $H_{0U} : \mu \geq \mu_0$ versus $H_{1U} : \mu < \mu_0$ is: reject H_{0U} if $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq t_U$. Thus, the union-intersection test of H_0 versus H_1 formed from these two LRTs is: reject H_0 if $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq t_L$ or $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq t_U$.

If $t_L = -t_U \geq 0$, the union-intersection test can be more simply expressed as

$$\textit{reject } H_0 \textit{ if } \frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} \geq t_L.$$

This test is called the two-sided t test.

The intersection-union method of test construction might be useful when the null hypothesis is conveniently expressed as a union. Suppose we wish to test the null hypothesis

$$H_0 : \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma.$$

Suppose that for each $\gamma \in \Gamma$, $\{\boldsymbol{x} : T_\gamma(\boldsymbol{x}) \in R_\gamma\}$ is the rejection region for a test of $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$. Then the rejection region for the test is

$$\bigcap_{\gamma \in \Gamma} \{\boldsymbol{x} : T_\gamma(\boldsymbol{x}) \in R_\gamma\}.$$

Example 8.2.5 *(***Acceptance sampling***) Two parameters that are important in assessing the quality upholstery fabric are θ_1, the mean breaking strength, and θ_2, the probability of passing a flammability test. Standards may dictate that θ_1 should be over 50 pounds and θ_2 should be over 0.95, and the fabric is acceptable only if it meets both of these standards. This can be modeled with the hypothesis test*

$$H_0 : \{\theta_1 \leq 50 \text{ or } \theta_2 \leq .95\} \quad \text{versus} \quad H_1 : \{\theta_1 > 50 \text{ and } \theta_2 > .95\},$$

where a batch of material is acceptable only if H_1 is acceptable.

Suppose $X_1, \ldots, X_n$ are measurements of breaking strength for n sample and are assumed to be iid $N(\theta_1, \sigma^2)$. The LRT of $H_{01} : \theta_1 \leq 50$ will reject H_{01} if $(\bar{X} - 50)/(S/\sqrt{n}) > t$. Suppose that we also have the results of m flammability test, denoted by $Y_1, \ldots, Y_m$, where $Y_i = 1$ if the i-th sample passes the test and $Y_i = 0$ otherwise. If $Y_1, \ldots, Y_m$ are modeled as iid Bernoulli(θ_2) random variables, the LRT will reject $H_{02} : \theta_2 \leq .95$ if $\sum_{i=1}^m Y_i > b$. Putting all of this together, the rejection region will be

$$\{(\boldsymbol{x}, \boldsymbol{y}) : \frac{\bar{x} - 50}{s/\sqrt{n}} > t \quad \text{and} \quad \sum_{i=1}^m y_i > b\}.$$

Table 8.1: Two types of errors in hypothesis testing

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct decision	Type I error
	H_1	Type II error	Correct decision

8.3 Methods of Evaluating Tests

8.3.1 Error Probabilities and the Power Function

A hypothesis test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ might make one of two types of errors, type I error and type II error. If $\theta \in \Theta_0$ but the hypothesis test incorrectly decides to reject H_0, then the test has made a type I error. If, on the other hand, $\theta \in \Theta_0^c$ but the test decides to] accept H_0, a type II error has been made. These two different situations are depicted in Table 8.1

Let R denote the rejection region for a test. Then

$$P_\theta(\boldsymbol{X} \in R) = \begin{cases} \text{probability of a type I error} & \text{if } \theta \in \Theta_0 \\ \text{one minus the probability of a type II error} & \text{if } \theta \in \Theta_0^c. \end{cases}$$

Definition 8.3.1 *The power function of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(\boldsymbol{X} \in R)$.*

Qualitatively, a good test has power function near 1 for most $\theta \in \Theta_0^c$ and near 0 for most $\theta \in \Theta_0$.

Example 8.3.1 *(***Binomial power function***) Let* $X \sim$ *binomial*$(5, \theta)$*. Consider testing* $H_0 : \theta \leq 1/2$ *versus* $H_1 : \theta > 1/2$*. Consider first the test that rejects* H_0 *if and only if all "successes" are observed. The power function for this test is*

$$\beta_1(\theta) = P_\theta(X \in R) = P_\theta(X = 5) = \theta^5.$$

In examining this power function, we might decide that although the probability of a type I error is acceptably low ($\beta_1(\theta) \leq (1/2)^5 = 0.0312$) *for all* $\theta \leq 1/2$*, the probability of a type II error is too high (*$\beta_1(\theta)$ *is too small) for most* $\theta > 1/2$*. To achieve smaller type II error probabilities, we might consider using the test that rejects* H_0 *if* $X = 3, 4,$ *or 5. The power function for this test is*

$$\beta_2(\theta) = P_\theta(X = 3, 4, \textit{or } 5) = \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta)^1 + \binom{5}{5}\theta^5(1-\theta)^0.$$

It is easy to see that the second test has achieved a smaller type II error probability in that $\beta_2(\theta)$ *is larger for* $\theta > 1/2$*. But the type I error probability is larger for the second test;* $\beta_2(\theta)$ *is larger for* $\theta \leq 1/2$*.*

Example 8.3.2 *(***Normal power function***)* *Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population, σ^2 known. An LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is a test that rejects H_0 if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > c$. The constant c can be any positive number. The power function of this test is*

$$\begin{aligned}\beta(\theta) &= P_\theta\big(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\big) = P_\theta\big(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\big) \\ &= P\big(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\big),\end{aligned}$$

where Z is a standard normal random variable. As θ increases from $-\infty$ to ∞, it is easy to see that

$$\lim_{\theta \to -\infty} \beta(\theta) = 0, \quad \lim_{\theta \to \infty} \beta(\theta) = 1, \quad \textit{and} \quad \beta(\theta_0) = \alpha \quad \textit{if } P(Z > c) = \alpha).$$

Suppose that the experimenter wishes to have a maximum Type I Error probability of 0.1, and a maximum type II error probability of 0.2 if $\theta \geq \theta_0 + \sigma$, that is,

$$\beta(\theta_0) = 0.1 \quad \textit{and} \quad \beta(\theta_0 + \sigma) = 0.8.$$

By choosing $c = 1.28$, we achieve $\beta(\theta_0) = P(Z > 1.28) = 0.1$, regardless of n. Now we wish to choose n so that $\beta(\theta_0 + \sigma) = P(Z > 1.28 - \sqrt{n}) = 0.8$. But, $P(Z > -0.84) = 0.8$. By setting $1.28 - \sqrt{n} = -0.84$ and solving for n yield $n = 4.49$. So choosing $c = 1.28$ and $n = 5$ yield a test with error probabilities controlled as specified by the experimenter.

For a fixed sample size, it is usually impossible to make both types of error probabilities arbitrarily small. In searching for a good test, it is common to restrict consideration to tests that control the type I error probability at a specified level. Within this class of tests we then search for tests that have type II error probability that is as small as possible.

Definition 8.3.2 *For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.*

Definition 8.3.3 *For $0 \leq \alpha \leq 1$, a test with power function $\beta(\alpha)$ is a level α test if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.*

Note that some authors do not make the distinction between the terms size and level that we have made, and sometimes these terms are used interchangeably. Experimenters commonly specify the level of the test they wish to use, with typical choices being $\alpha = 0.01, 0.05$, and 0.1. Be aware that, in fixing the level of the test, the experimenter is controlling only the type I error probabilities, not the type II error.

Example 8.3.3 *(**Size of LRT**) In general, a size α LRT is constructed by choosing c such that $\sup_{\theta \in \Theta_0} P_\theta(\lambda(\boldsymbol{X}) \leq c) = \alpha$. How that c is determined depends on the particular problem. For example, in Example 8.2.1, Θ_0 consists of the single point $\theta = \theta_0$ and $\sqrt{n}(\bar{X} - \theta_0) \sim N(0, 1)$ if $\theta = \theta_0$. So the test*

$$\textit{reject } H_0 \textit{ if } |\bar{X} - \theta_0| \geq z_{\alpha/2}/\sqrt{n},$$

where $z_{\alpha/2}$ satisfies $P(Z > z_{\alpha/2}) = \alpha/2$ with $Z \sim N(0, 1)$, is the size α LRT.

Specifically, this corresponds to choosing $c = \exp(-z^2_{\alpha/2})$*, this this is not an important point.*

Example 8.3.4 *(***Size of union-intersection test***) The problem of finding a size α union-intersection test in Example 8.2.4 involves finding constant t_L and t_U such that*

$$\sup_{\theta\in\Theta_0} P_\theta\big(\frac{\bar{X}-\mu_0}{\sqrt{S^2/\sqrt{n}}} \geq t_L \quad or \quad \frac{\bar{X}-\mu_0}{\sqrt{S^2/\sqrt{n}}} \leq t_U\big) = \alpha.$$

But under H_0, $\frac{\bar{X}-\mu_0}{\sqrt{S^2/\sqrt{n}}}$ has a student's t distribution with $n-1$ degrees of freedom. So any choice of $t_U = t_{n-1,1-\alpha_1}$ and $t_L = t_{n-1,\alpha_2}$, with $\alpha_1+\alpha_2=\alpha$, will yield a test with type I error probability of exactly α for all $\theta\in\Theta_0$. The usual choice is $t_L = -t_U = t_{n-1,\alpha/2}$.

Other than α level, there are other features of a test that might also be of concern. For example, we would like a test to be more likely to reject H_0 if $\theta\in\Theta_0^c$ than if $\theta\in\Theta_0$.

Definition 8.3.4 *A test with power function $\beta(\theta)$ is unbiased if $\beta(\theta') \geq \beta(\theta'')$ for every $\theta'\in\Theta_0^c$ and $\theta''\in\Theta_0$.*

Example 8.3.5 *(***Continuation of Example 8.3.2***) An LRT of $H_0:\theta\leq\theta_0$ versus $H_1:\theta>\theta_0$ has power function*

$$\beta(\theta) = P\big(Z > c + \frac{\theta_0-\theta}{\sigma/sqrtn}\big),$$

where $Z\sim N(0,1)$. Since $\beta(\theta)$ is an increasing function of θ (for fixed θ_0), it follows that

$$\beta(\theta) > \beta(\theta_0) = \max_{t\leq\theta_0}\beta(t) \quad for\ all\ \theta>\theta_0$$

and, hence, that the test is unbiased.

8.3.2 Most Powerful Tests

Definition 8.3.5 *Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class $\mathcal{C}$, with power function $\beta(\theta)$, is a uniformly most powerful (UMP) class $\mathcal{C}$ test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^c$ and every $\beta'(\theta)$ that is a power function of a test in class $\mathcal{C}$.*

In this section, the class $\mathcal{C}$ will be the class of all level α tests. The test described in the above definition is then called a UMP level α test. The requirements in the definition are so strong that UMP tests do not exist in many realistic problems. But in problems that have UMP tests, a UMP test might well be considered the best test in the class. The following famous theorem clearly describes which test are UMP tests if they exist.

Theorem 8.3.1 *(***Neyman-Pearson Lemma***) Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(\boldsymbol{x}|\theta_i)$, $i = 0, 1$, using a test with rejection region R that satisfies*

$$\begin{aligned} x \in R \quad &\text{if} \quad f(\boldsymbol{x}|\theta_1) > k f(\boldsymbol{x}|\theta_0) \\ \textit{and} \qquad & \qquad (8.1) \\ x \in R^c \quad &\text{if} \quad f(\boldsymbol{x}|\theta_1) < k f(\boldsymbol{x}|\theta_0) \end{aligned}$$

for some $k \geq 0$, and

$$\alpha = P_{\theta_0}(\boldsymbol{X} \in R). \tag{8.2}$$

Then

a. *(Sufficiency) Any test that satisfies (8.1) and (8.2) is a UMP level α test.*

b. (Necessity) If there exists a test satisfying (8.1) and (8.2) with $k > 0$, then every UMP level α test is a size α test (satisfying (8.2)) and every UMP level α test satisfies (8.1) except perhaps on a set A satisfying $P_{\theta_0}(\boldsymbol{X} \in A) = P_{\theta_1}(\boldsymbol{X} \in A) = 0$.

PROOF: We will prove the theorem for the case that $f(\boldsymbol{x}|\theta_0)$ and $f(\boldsymbol{x}|\theta_1)$ are pdfs of continuous random variables. The proof for discrete random variables can be accomplished by replacing integrals with sums. Note first that any test satisfying (8.2) is a size α and, hence, a level α test because $\sup_{\theta\in\Theta_0} P_\theta(\boldsymbol{X} \in R) = P_{\theta_0}(\boldsymbol{X} \in R) = \alpha$, since Θ_0 has only one point.

To ease notation, we define a test function, a function on the sample space that is 1 if $\boldsymbol{x} \in R$ and 0 if
$bx \in R^c$. That is, it is the indicator function of the rejection region. Let $\phi(\boldsymbol{x})$ be the test function of a test satisfying (8.1) and (8.2). Let $\phi'(\boldsymbol{x})$ be the test function of any other level α test, and let $\beta(\theta)$ and $\beta'(\theta)$ be the power functions corresponding to the tests ϕ and ϕ', respectively. Because $0 \le \phi'(\boldsymbol{x}) \le 1$, (8.1) implies that

$$(\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x}))(f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0)) \ge 0$$

for every $\boldsymbol{x}$ (since $\phi = 1$ if $f(\boldsymbol{x}|\theta_1) > kf(\boldsymbol{x}|\theta_0)$ and $\phi = 0$ if $f(\boldsymbol{x}|\theta_1) < kf(\boldsymbol{x}|\theta_0)$). Thus

$$0 \le \int (\phi(\boldsymbol{x})-\phi'(\boldsymbol{x}))(f(\boldsymbol{x}|\theta_1)-kf(\boldsymbol{x}|\theta_0))dx = \beta(\theta_1)-\beta'(\theta_1)-k(\beta(\theta_0)-\beta(\theta_1)). \tag{8.3}$$

Statement (a) is proved by noting that, since ϕ' is a level α test and ϕ is a

size α test, $\beta(\theta_0) - \beta'(\theta_0) = \alpha - \beta'(\theta_0) \geq 0$. Thus (8.3) and $k \geq 0$ imply that

$$0 \leq \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta(\theta_1)) \leq \beta(\theta_1) - \beta'(\theta_1),$$

showing that $\beta(\theta) \geq \beta'(\theta)$ and hence ϕ has greater power than ϕ'. Since ϕ' was an arbitrary level α test and θ_1 is the only point in Θ_0^c, ϕ is a UMP level α test.

To prove statement (b), let ϕ' now be the test function for any UMP level α test. By part (a), ϕ, the test satisfying (8.1) and (8.2), is also a UMP level α test, thus $\beta(\theta_1) = \beta'(\theta_1)$. This fact, (8.3), and $k > 0$ imply

$$\alpha - \beta'(\theta_0) = \beta(\theta_0) - \beta'(\theta_0) \leq 0.$$

Now, since ϕ' is a level α test, $\beta'(\theta) \leq \alpha$. Thus $\beta'(\theta_0) = \alpha$, that is, ϕ' is a size α test, and this also implies that (8.3) is an equality in this case. But the non-negative integrand $(\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x}))(f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0))$ will have a zero integral only if ϕ' satisfies (8.1), except perhaps on a set A with $\int_A f(\boldsymbol{x}|\theta_i)d\boldsymbol{x} = 0$. This implies that the last assertion in statement (b) is true. □

The following corollary connects the Neyman-Pearson Lemma to sufficiency.

Corollary 8.3.1 *Consider the hypothesis problem posed in Theorem 8.3.1. Suppose $T(\boldsymbol{X})$ is a sufficient statistic for θ and $g(t|\theta_i)$ is the pdf or pmf of T corresponding to θ_i, $i = 0, 1$. Then any test based on T with rejection region S (a subset of the sample space of T) is a UMP level α test if it satisfies*

$$\begin{gathered} t \in S \quad \text{if} \quad g(t|\theta_1) > kg(t|\theta_0) \\ \text{and} \\ t \in S^c \quad \text{if} \quad g(t|\theta_1) < kg(t|\theta_0) \end{gathered} \tag{8.4}$$

for some $k \geq 0$, where

$$\alpha = P_{\theta_0}(T \in S). \tag{8.5}$$

PROOF: In terms of the original sample $\boldsymbol{X}$, the test based on T has the rejection region $R = \{\boldsymbol{x} : T(\boldsymbol{x}) \in S\}$. By the Factorization Theorem, the pdf or pmf of $\boldsymbol{X}$ can be written as $f(\boldsymbol{x}|\theta_i) = g(T(\boldsymbol{x})|\theta_i)h(\boldsymbol{x})$, $i = 0, 1$, for some nonnegative function $h(\boldsymbol{x})$. Multiplying the inequalities in (8.4) by this nonnegative function, we see that R satisfies

$$x \in R \quad \text{if} \quad f(\boldsymbol{x}|\theta_1) = g(T(\boldsymbol{x})|\theta_1)h(\boldsymbol{x}) > kg(T(\boldsymbol{x})|\theta_0)h(\boldsymbol{x}) = kf(\boldsymbol{x}|\theta_0)$$

and

$$x \in R^c \quad \text{if} \quad f(\boldsymbol{x}|\theta_1) = g(T(\boldsymbol{x})|theta_1)h(\boldsymbol{x}) < kg(T(\boldsymbol{x})|\theta_0)h(\boldsymbol{x}) = kf(\boldsymbol{x}|\theta_0)$$

Also, by (8.5),

$$P_{\theta_0}(\boldsymbol{X} \in R) = P_{\theta_0}(T(\boldsymbol{X}) \in S) = \alpha.$$

So, by the sufficiency part of the Neyman-Pearson Lemma, the test based on T is a UMP level α test. □

Example 8.3.6 *(***UMP binomial test***) Let* $X \sim$ *Binomial*$(2, \theta)$*. We want to test* $H_0 : \theta = \frac{1}{2}$ *versus* $H_1 : \theta = \frac{3}{4}$*. Calculating the ratios of the pmfs gives*

$$\frac{f(0|\theta = 3/4)}{f(0|\theta = 1/2)} = \frac{1}{4}, \quad \frac{f(1|\theta = 3/4)}{f(1|\theta = 1/2)} = \frac{3}{4}, \quad \textit{and} \quad \frac{f(2|\theta = 3/4)}{f(2|\theta = 1/2)} = \frac{9}{4}.$$

If we choose $3/4 < k < 9/4$*, the Neyman-Pearson Lemma says that the test that rejects* H_0 *if* $X = 2$ *is the UMP level* $\alpha = P(X = 2|\theta = 1/2) = \frac{1}{4}$ *test. If we choose* $1/4 < k < 3/4$*, the Neyman-Pearson Lemma says that the test that rejects* H_0 *if* $X = 1$ *or 2 is the UMP level* $\alpha = P(X = 1 \text{ or } 2|\theta = 1/2) = 3/4$ *test. Choosing* $k < 1/4$ *or* $k > 9/4$ *yields the UMP level* $\alpha = 1$ *or level* $\alpha = 0$ *test.*

Note that if $k = 3/4$*, then (8.1) says we must reject* H_0 *for the sample point* $x = 2$ *and accept* H_0 *for* $x = 0$ *but leaves our action for* $x = 1$ *undetermined. But if we accept* H_0 *for* $x = 1$*, we get the UMP level* $\alpha = 1/4$ *test as above. If we reject* H_0 *for* $x = 1$*, we get the UMP level* $\alpha = 3/4$ *test as above.*

Example 8.3.7 *(***UMP normal test***) Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population, σ^2 known. The sample mean $\bar{X}$ is a sufficient statistic for θ. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$. The inequality (8.4), $g(\bar{x}|\theta_1) > kg(\bar{x}|\theta_0)$, is equivalent to*

$$\bar{x} < \frac{(2\sigma^2 \log k)/n - \theta_0^2 + \theta_1^2}{2(\theta_1 - \theta_0)}.$$

Note that the right-hand side increases from $-\infty$ to ∞ as k increases from 0 to ∞. Thus, by Corollary 8.3.1, the test with rejection region $\bar{x} < c$ is the UMP level α test, where $\alpha = P_{\theta_0}(\bar{X} < c)$. If a particular α is specified, then the UMP test rejects H_0 if $\bar{X} < c = -\sigma z_\alpha/\sqrt{n} + \theta_0$. The choice of c ensures that (8.5) is true.

Hypotheses, such as H_0 and H_1 in the Neyman-Pearson Lemma, that specify only one possible distribution for the sample $\boldsymbol{X}$ are called *simple hypotheses*. In most realistic problems, the hypotheses of interest specify more than one possible distribution for the sample. Such hypotheses are called *composite hypotheses*. Since Definition 8.3.5 requires a UMP test to be most powerful against each individual $\theta \in \Theta_0^c$, the Neyman-Pearson Lemma can be used to find UMP tests in problems involving composite hypotheses. In particular, hypotheses that assert that a univariate parameter is large, for example, $H : \theta \geq \theta_0$, or small, for example, $H : \theta < \theta_0$, are called *one-sided hypotheses*. Hypotheses that assert that a parameter is either large or small, for example, $H : \theta \neq \theta_0$, are called *two-sided hypotheses*. A large class of problems that admit UMP level α tests involve one-sided hypotheses and pdfs or pmfs with the monotone likelihood ratio property.

Definition 8.3.6 *A family of pdfs or pmfs $\{g(t|\theta) : \theta \in \Theta\}$ for a univariate random variable T with real-valued parameter θ has a monotone likelihood ratio (MLR) if, for every $\theta_2 > \theta_1$, $g(t|\theta_2)/g(t|\theta_1)$ is a monotone (nonincreasing or nondecreasing) function of t on $\{t : g(t|\theta_1) > 0$ or $g(t|\theta_2) > 0\}$. Note that $c/0$ is defined as ∞ if $c > 0$.*

Many common families of distributions have an MLR. For example, the normal (known variance, unknown mean), Poisson, and binomial all have an MLR. Indeed, any regular exponential family with $g(t|\theta) = h(t)c(\theta)e^{w(\theta)t}$ has an MLR if $w(\theta)$ is a nondecreasing function.

Theorem 8.3.2 *(***Karlin-Rubin***) Consider testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of pdfs or pmfs $\{g(t|\theta) : \theta \in \Theta\}$ of T has an MLR. Then for any t_0, the test that rejects H_0 if and only if $T > t_0$ is a UMP level α test, where $\alpha = P_{\theta_0}(T > t_0)$.*

PROOF: Let $\beta(\theta) = P_\theta(T > t_0)$ be the power function of the test. We first show that $\beta(\theta)$ is nondecreasing. For $\theta_2 > \theta_1$, we have

$$\frac{d}{dt_0}[\beta(\theta_1) - \beta(\theta_2)] = -g(t_0|\theta_1) + g(t_0|\theta_2) = g(t_0|\theta_1)\big(\frac{g(t_0|\theta_2)}{g(t_0|\theta_1)} - 1\big).$$

Because g has MLR, the ratio on the right-hand side is nondecreasing, so the derivative can only change sign from negative to positive showing that any interior extremum is a minimum. Thus $\beta(\theta_1) - \beta(\theta_2)$ is maximized by its value at ∞ or $-\infty$, which is zero.

Fix $\theta' > \theta_0$ and consider testing $H_0' : \theta = \theta_0$ versus $H_0' : \theta = \theta'$. Since $\beta(\theta)$ is nondecreasing, so

i. $\sup_{\theta \le \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$, and this is a level α test.

ii. If we define

$$k' = \inf_{t \in \mathcal{T}} \frac{g(t|\theta')}{g(t|\theta_0)},$$

where $\mathcal{T} = \{t : t > t_0 \text{ and either } g(t|\theta') > 0 \text{ or } g(t|\theta_0) > 0\}$, it follows that

$$T > t_0 \Leftrightarrow \frac{g(t|\theta')}{g(t|\theta_0)} > k'.$$

Together with Corollary 8.3.1, (i) and (ii) imply that $\beta(\theta') \ge \beta^*(\theta')$, where $\beta^*(\theta)$ is the power function for any other level α test of H_0', that is, any test satisfying $\beta(\theta_0) \le \alpha$. However, any level α test of H_0 satisfies $\beta^*(\theta_0) \le$

$\sup_{\theta\in\Theta_0}\beta^*(\theta) \le \alpha$. Thus, $\beta(\theta') \ge \beta^*(\theta')$ for any level α test of H_0. Since θ' was arbitrary, the test is a UMP level α test. □

By an analogous argument, it can be shown that under the conditions of Theorem 8.3.2, the test that rejects $H_0 : \theta \ge \theta_0$ in favor of $H_1 : \theta < \theta_0$ if and only if $T < t_0$ is a UMP level $\alpha = P_{\theta_0}(T < t_0)$ test.

Example 8.3.8 *(***Continuation of Example 8.3.7***) Consider testing* $H_0' : \theta \ge \theta_0$ *versus* $H_1' : \theta < \theta_0$ *using the test that rejects* H_0' *if*

$$\bar{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0.$$

As $\bar{X}$ *is sufficient and its distribution has an MLR, it follows from Theorem 8.3.2 that the test is a UMP level* α *test in this problem. The power function of this test,*

$$\beta(\theta) = P_\theta\big(\bar{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\big),$$

is a decreasing function of θ*, the value of* α *is given by* $\sup_{\theta\ge\theta_0}\beta(\theta) = \beta(\theta_0) = \alpha$.

Although most experimenters would choose to use a UMP level α test if they knew of one, unfortunately, for many problems there is no UMP level α test. That is, no UMP test exists because the class of level α tests is so large that no one test dominates all the others in terms of power. In such cases, a common method of continuing the search for a good test is to consider some subset of the class of level α tests and attempt to find a UMP test in this subset.

Example 8.3.9 *(***Nonexistence of UMP test***) Let $X_1, \ldots, X_n$ be iid $N(\theta, \sigma^2)$, σ^2 known. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. For a specified value of α, a level α test in this problem is any test that satisfies*

$$P_{\theta_0}(\textit{reject } H_0) \leq \alpha. \tag{8.6}$$

Consider an alternative parameter point $\theta_1 < \theta_0$. The analysis in Example 8.3.8 shows that, among all tests that satisfy (8.6), the test that rejects H_0 if $\bar{X} < -\sigma z_\alpha/\sqrt{n} + \theta_0$ has the highest possible power at θ_1. Call this Test 1. Furthermore, by part (b) (necessity) of the Neyman-Pearson Lemma, any other level α test that has as high a power as Test 1 at θ_1 must have the same rejection region as Test 1 except possibly for a set A satisfying $\int_A f(\boldsymbol{x}|\theta_i)d\boldsymbol{x} = 0$. Thus, if a UMP level α test exists for this problem, it must be Test 1 because no other test has as high a power as Test 1 at θ_1.

Now consider Test 2, which rejects H_0 if $\bar{X} > \sigma z_\alpha/\sqrt{n} + \theta_0$. Test 2 is also a level α test. Let $\beta_i(\theta)$ denote the power function of Test i. For any $\theta_2 > \theta_0$,

$$\begin{aligned}
\beta_2(\theta_2) &= P_{\theta_2}\big(\bar{X} > \frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\big) == P_{\theta_2}\big(\frac{\bar{X} - \theta_2}{\sigma/\sqrt{n}} > z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}\big) \\
&> P(Z > z_\alpha) = P(Z < -z_\alpha) \\
&> P_{\theta_2}\big(\frac{\bar{X} - \theta_2}{\sigma/\sqrt{n}} < -z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}\big) = P_{\theta_2}\big(\bar{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\big) = \beta_1(\theta_2).
\end{aligned}$$

Thus Test 1 is not a UMP level α test because Test 2 has a higher power than Test 1 at θ_2. Earlier we showed that if there were a UMP level α test, it would have to be Test 1. Therefore, no UMP level α test exists in this problem.

When no UMP level α test exists within the class of all tests, we might try to find a UMP level α test within the class of unbiased tests. Consider Test

3, which rejects $H_0 : \theta = \theta_0$ *in favor of* $H_1 : \theta \neq \theta_0$ *if and only if*

$$\bar{X} > \sigma z_{\alpha/2}/\sqrt{n} + \theta_0 \quad or \quad \bar{X} < -\sigma z_{\alpha/2}/sqrtn + \theta_0.$$

This test is actually a UMP unbiased level α *test; that is, it is UMP in the class of unbiased tests. The following figure shows the power function* $\beta_1(\theta)$*,* $\beta_2(\theta)$ *and* $\beta_3(\theta)$*. If our interest is in rejecting* H_0 *for both large and small values of* θ*, the figure shows that Test 3 is better overall than either Test 1 or Test 2.*

8.3.3 Sizes of Union-Intersection and Intersection-Union tests

Because of the simple way in which they are constructed, the sizes of union-intersection tests (UIT) and intersection-union tests (IUT) can often be bounded above by the sizes of some other tests. Such bounds are useful if a level α test is wanted, but the size if the UIT or IUT is too difficult to evaluate. In this section we discuss these bounds and give examples in which the bounds are sharp, that is, the size of the test is equal to the bound.

Theorem 8.3.3 *Consider a UIT test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$, where $\Theta_0 = \bigcap_{\gamma\in\Gamma} \Theta_\gamma$. Let $\lambda_\gamma(\boldsymbol{x})$ be the LRT statistic for testing $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$, and let $\lambda(\boldsymbol{x})$ be the LRT statistic for $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. Define $T(\boldsymbol{x}) = \inf_{\gamma\in\Gamma} \lambda_\gamma(\boldsymbol{x})$, and form the UIT with rejection region*

$$\{\boldsymbol{x} : \lambda_\gamma(\boldsymbol{x}) < c \text{ for some } \gamma \in \Gamma\} = \{\boldsymbol{x} : T(\boldsymbol{x}) < c\}.$$

Also consider the usual LRT with rejection region $\{\boldsymbol{x} : \lambda_\gamma(\boldsymbol{x}) < c\}$. Then

a. $T(\boldsymbol{x}) > \lambda(\boldsymbol{x})$ for every $\boldsymbol{x}$.

b. If $\beta_T(\theta)$ and $\beta_\lambda(\theta)$ are the power functions for the tests based on T and λ, respectively, then $\beta_T(\theta) \le \beta_\lambda(\theta)$ for every $\theta \in \Theta$.

c. If the LRT is a level α test, then the UIT is a level α test.

PROOF: Since $\Theta_0 = \bigcap_{\gamma\in\Gamma} \Theta_\gamma \subset \Theta_\gamma$ for any γ, from Definition 8.2.1 we see that for any $\boldsymbol{x}$,

$$\lambda_\gamma(\boldsymbol{x}) \ge \lambda(\boldsymbol{x}) \quad \text{for each } \gamma \in \Gamma$$

because the region of maximization is bigger for the individual λ_γ. Thus $T(\boldsymbol{x}) = \inf_{\gamma\in\Gamma} \lambda_\gamma(\boldsymbol{x}) \geq \lambda(\boldsymbol{x})$, proving (a).

By (a), $\{\boldsymbol{x} : T(\boldsymbol{x}) < c\} \subset \{\boldsymbol{x} : \lambda(\boldsymbol{x}) < c\}$, so

$$\beta_T(\theta) = P_\theta(T(\boldsymbol{X}) < c) \leq P_\theta(\lambda(\boldsymbol{x}) < c) = \beta_\lambda(\theta),$$

proving (b).

Since (b) holds for every θ, $\sup_{\theta\in\Theta_0} \beta_T(\theta) \leq \sup_{\theta\in\Theta_0} \beta_\lambda(\theta) \leq \alpha$, proving (c). □

Since the LRT is uniformly more powerful than the UIT in Theorem 8.3.3, we might ask why we should use the UIT. One reason is that the UIT has a smaller type I error probability for every $\theta \in \Theta_0$. Furthermore, if H_0 is rejected, we may wish to look at the individual tests of $H_{0\gamma}$ to see why. The possibility of gaining additional information by looking at the $H_{0\gamma}$ individually, rather than looking only at the overall LRT, is evident.

Now we investigate the sizes of IUTs. Recall that in an IUT, we test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$, where $\Theta_0 = \bigcap_{\gamma\in\Gamma} \Theta_\gamma$. An IUT has a rejection region of the form $R = \bigcap_{\gamma\in\Gamma} R_\gamma$, where R_γ is the rejection region for a test of $H_{0\gamma} : \theta \in \Theta_\gamma$.

Theorem 8.3.4 *Let α_γ be the size of the test of $H_{0\gamma}$ with rejection region R_γ. Then the IUT with rejection region $R = \bigcap_{\gamma\in\Gamma} R_\gamma$ is a level $\alpha = \sup_{\gamma\in\Gamma} \alpha_\gamma$ test.*

PROOF: Let $\theta \in \Theta_0$. Then $\theta \in \Theta_\gamma$ for some γ and

$$P_\theta(\boldsymbol{x} \in R) \le P_\theta(\boldsymbol{X} \in R_\gamma) \le \alpha_\gamma \le \alpha.$$

Since $\theta \in \Theta_0$ was arbitrary, the IUT is a level α test. □

Note that Theorem 8.3.3 applies only to UITs constructed from likelihood ratio tests. In contrast, Theorem 8.3.4 applies to any IUT. The bound in Theorem 8.3.3 is the size of the LRT, which, in a complicated problem, may be difficult to compute. In Theorem 8.3.4, however, the LRT need not be used to obtain the upper bound. Any test of $H_{0\gamma}$ with known size α_γ can be used, and then the upper bound on the size of the IUT is given in terms of the known sizes α_γ, $\gamma \in \Gamma$.

The IUT in Theorem 8.3.4 is a level α test. But the size of the IUT may be much less than α; the IUT may be very conservative. The following theorem gives conditions under which the sizes of the IUT is exactly α and the IUT is not too conservative.

Theorem 8.3.5 *Consider testing $H_0 : \theta \in \bigcup_{j=1}^{k} \Theta_j$, where k is a finite positive integer. For each $j = 1, \ldots, k$, let R_j be the rejection region of a level α test of H_{0j}. Suppose that for some $i = 1, \ldots, k$, there exists a sequence of parameter points, $\theta_l \in \Theta_i$, $l = 1, 2, \ldots$, such that*

i. $\lim_{l\to\infty} P_{\theta_l}(\boldsymbol{X} \in R_i) = \alpha$,

ii. for each $j = 1, \ldots, k$, $j \neq i$, $\lim_{l\to\infty} P_{\theta_l}(\boldsymbol{X} \in R_j) = 1$.

Then, the IUT with rejection region $R = \bigcap_{j=1}^{k} R_j$ is a size α test.

PROOF: By Theorem 8.3.4, R is a level α test, that is,

$$\sup_{\theta\in\Theta_0} P_\theta(\boldsymbol{X} \in R) \leq \alpha. \tag{8.7}$$

But, because all the parameter points θ_l satisfy $\theta_l \in \Theta_i \subset \Theta_0$,

$$\begin{aligned}
\sup_{\theta\in\Theta_0} P_\theta(\boldsymbol{X} \in R) &\geq \lim_{l\to\infty} P_{\theta_l}(\boldsymbol{X} \in R) = \lim_{l\to\infty} P_{\theta_l}(\boldsymbol{X} \in \bigcap_{j=1}^{k} R_j) \\
&\geq \lim_{l\to\infty} \sum_{j=1}^{k} P_\theta(\boldsymbol{X} \in R_j) - (k-1) \qquad \text{(Bonferroni's Inequality)} \\
&= (k-1) + \alpha - (k-1) = \alpha.
\end{aligned}$$

This and (8.7) imply the test has size exactly equal to α. □

Example 8.3.10 *(***Intersection-union test***)* *In Example (8.2.5), let* $n = m = 58$*,* $t = 1.672$*, and* $b = 57$*. The each of the individual tests has size* $\alpha = 0.05$ *(approximately). Therefore, by Theorem 8.3.4, the IUT is a level* $\alpha = 0.05$ *test. In fact, this test is a size* $\alpha = 0.05$ *test. To see this consider a sequence of parameter points* $\theta_l = (\theta_{1l}, \theta_2)$*, with* $\theta_{1l} \to \infty$ *and* $\theta_2 = 0.95$*. All such parameter points are in* Θ_0 *because* $\theta_2 \leq 0.95$*. Also,* $P_{\theta_l}(\boldsymbol{X} \in R_1) \to 1$ *as* $\theta_{1l} \to \infty$*, while* $P_{\theta_l}(\boldsymbol{X} \in R_2) = 0.05$ *for all* l *because* $\theta_2 = 0.95$*. Thus, by Theorem 8.3.5, the IUT is a size* α *test.*

Note that, in Example 8.3.25, only the marginal distributions of the $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ were used to find the size of the test. This point is extremely important and directly relates to the usefulness of IUTs, because the joint distribution is often difficult to know and, if known, often difficult to work with.

8.3.4 p-Values

Definition 8.3.7 *A p-value $p(\boldsymbol{X})$ is a test statistic satisfying $0 \leq p(\boldsymbol{x}) \leq 1$ for every sample point $\boldsymbol{x}$. Small values of $p(\boldsymbol{X})$ gives evidence that H_1 is true. A p-value is valid if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,*

$$P_\theta(p(\boldsymbol{x}) \leq \alpha) \leq \alpha.$$

If $p(\boldsymbol{X})$ is a valid p-value, it is easy to construct a level α test based on $p(\boldsymbol{X})$. The test that rejects H_0 if and only if $p(\boldsymbol{X}) \leq \alpha$ is a level α test. An advantage to reporting a test result via a p-value is that each reader can choose the α he or she considers appropriate and then can compare the reported $p(\boldsymbol{x})$ to α and know whether these data lead to acceptance or rejection of H_0. Furthermore, the smaller the p-value, the stronger the evidence for rejecting H_0. Hence, a p-value reports the results of a test on a more continuous scale, rather than just the dichotomous decision "Accept H_0" or "Reject H_0". The following theorem gives a common way how to define a valid p-value.

Theorem 8.3.6 *let $W(\boldsymbol{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\boldsymbol{x}$, define*

$$p(\boldsymbol{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\boldsymbol{X}) \geq W(\boldsymbol{x})). \tag{8.8}$$

Then, $p(\boldsymbol{X})$ is a valid p-value.

PROOF: Fix $\theta \in \Theta_0$. Let $F_\theta(w)$ denote the cdf of $-W(\boldsymbol{X})$. Define

$$p_\theta(\boldsymbol{x}) = P_\theta(W(\boldsymbol{X}) \geq W(\boldsymbol{x})) = P_\theta(-W(\boldsymbol{X}) \leq -W(\boldsymbol{x})) = F_\theta(-W(\boldsymbol{x})).$$

Then the random variable $p_\theta(\boldsymbol{X})$ is equal to $F_\theta(-W(\boldsymbol{x}))$. Hence, by Probability Integral Transformation, the distribution of $p_\theta(\boldsymbol{X})$ is stochastically greater than or equal to a uniform(0,1) distribution. That is, for every $0 \leq \alpha \leq 1$, $P_\theta(p_\theta(\boldsymbol{X}) \leq \alpha) \leq \alpha$. Because $p(\boldsymbol{x}) = \sup_{\theta' \in \Theta_0} p_{\theta'}(\boldsymbol{x}) \geq p_\theta(\boldsymbol{x})$ for every $\boldsymbol{x}$,

$$P_\theta(p(\boldsymbol{X}) \leq \alpha) \leq P_\theta(p_\theta(\boldsymbol{X}) \leq \alpha) \leq \alpha.$$

This is true for every $\theta \in \Theta_0$ and for every $0 \leq \alpha \leq 1$, $p(\boldsymbol{X})$ is a valid p-value. □

Example 8.3.11 *(***Two-sided normal** p**-value**$)$ *Let* $X_1, \ldots, X_n$ *be a random sample from a* $N(\mu, \sigma^2)$ *population. Consider testing* $H_0 : \mu = \mu_0$ *versus* $H_1 : \mu \neq \mu_0$. *By Exercise 8.3.4, the LRT rejects* H_0 *for large values of* $W(\boldsymbol{X}) = |\bar{X} - \mu_0|/(S/\sqrt{n})$. *If* $\mu = \mu_0$, *regardless of the value of* σ, $(\bar{X} - \mu_0)/(S/\sqrt{n})$ *has a Student's t distribution with* $n - 1$ *degrees of freedom. Thus, in calculating (8.8), the probability is the same for all values of* θ, *,that is, all values of* σ. *Thus, the p-value from (8.8) for this two-sided t test is* $p(\boldsymbol{x}) = 2P(T_{n-1} \geq |\bar{x} - \mu_0|/(s/\sqrt{n}))$, *where* T_{n-1} *has a Student's t distribution with* $n - 1$ *degrees of freedom.*

Example 8.3.12 *(***One-sided normal** *p***-value***) Again consider the normal model of Example 8.3.11, but consider testing* $H_0 : \mu \leq \mu_0$ *versus* $H_1 : \mu > \mu_0$*. By Exercise 8.3.3, the LRT rejects* H_0 *for large values of* $W(\boldsymbol{X}) = (\bar{X} - \mu_0)/(S/\sqrt{n})$*. The following arguments shows that, for this statistic, the supremum in (8.8) always occurs at a parameter* (μ_0, σ)*, and the value of* σ *used does not matter. Consider any* $\mu \leq \mu_0$ *and any* σ*:*

$$
\begin{aligned}
&P_{\mu,\sigma}(W(\boldsymbol{X}) \geq W(\boldsymbol{x})) = P_{\mu,\sigma}\big(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq W(\boldsymbol{x})\big) \\
&= P_{\mu,\sigma}\big(\frac{\bar{X} - \mu}{S/\sqrt{n}} \geq W(\boldsymbol{x}) + \frac{\mu_0 - \mu}{S/sqrtn}\big) = P_{\mu,\sigma}\big(T_{n-1} \geq W(\boldsymbol{x}) + \frac{\mu_0 - \mu}{S/sqrtn}\big) \\
&\leq P(T_{n-1} \geq W(\boldsymbol{x})).
\end{aligned}
$$

Here again, T_{n-1} *has a Student's t distribution with* $n-1$ *degrees of freedom. The inequality in the last line is true because* $\mu_0 \geq \mu$*. The subscript on* P *is dropped here, because this probability does not depend on* (μ, σ)*. Furthermore,*

$$
P(T_{n-1} \geq W(\boldsymbol{x})) = P_{\mu_0,\sigma}\big(\frac{\bar{X} - \mu_0}{S/sqrtn} \geq W(\boldsymbol{x})\big) == P_{\mu_0,\sigma}\big(W(\boldsymbol{X}) \geq W(\boldsymbol{x})\big),
$$

and this probability is one of those considered in the calculation of the supremum in (8.8) because $(\mu_0, \sigma) \in \Theta_0$*. Thus, the p-value for this one-sided t test is* $p(\boldsymbol{x}) = P(T_{n-1} \geq W(\boldsymbol{x})) = P(T_{n-1} \geq (\bar{x} - \mu_0)/(s/\sqrt{n}))$*.*

Another method for defining a valid p-value involves conditioning on a sufficient statistic. Suppose $S(\boldsymbol{X})$ is a sufficient statistic for the model $\{f(\boldsymbol{x}|\theta) : \theta \in \Theta_0\}$. If the null hypothesis is true, the conditional distribution of $\boldsymbol{X}$ given $S = s$ does not depend on θ. Again, let $W(\boldsymbol{X})$ denote a test statistic for which large values give evidence that H_1 is true. Then, for

each sample point $\boldsymbol{X}$ define

$$p(\boldsymbol{x}) = P(W(\boldsymbol{X}) \geq W(\boldsymbol{x})|S = S(\boldsymbol{x})). \tag{8.9}$$

Arguing as in Theorem 8.3.6, but considering only the single distribution that is the conditional distribution of $\boldsymbol{X}$ given $S = s$, we see that, for any $0 \leq \alpha \leq 1$,

$$P(p(\boldsymbol{X}) \leq \alpha|S = s) \leq \alpha.$$

Then, for any $\theta \in \Theta_0$, unconditionally we have

$$P_\theta(p(\boldsymbol{X}) \leq \alpha) = \sum_s P(p(\boldsymbol{X}) \leq \alpha|S = s)P_\theta(S = s) \leq \sum_s \alpha P_\theta(S = s) \leq \alpha.$$

Thus, $p(\boldsymbol{X})$ defined in (8.9) is a valid p-value.

Example 8.3.13 *(***Fisher's Exact Test***) Let S_1 and S_2 be independent observations with $S_1 \sim$ Binomial(n_1, p) and $S_2 \sim$ Binomial(n_2, p). Consider testing $H_0 : p_1 = p_2$ versus $H_1 : p_1 > p_2$. Under H_0, if we let p denote the common value of $p_1 = p_2$, the joint pmf of (S_1, S_2) is*

$$f(s1, s2|p) = \binom{n_1}{s_1}\binom{n_2}{s_2} p^{s_1+s_2}(1-p)^{n_1+n_2-(s_1+s_2)}.$$

Thus, $S_1 + S_2$ is a sufficient statistic under H_0. Given the value of $S = s$, it is reasonable to use S_1 as a test statistic and reject H_0 in favor of H_1 for large values of S_1, because large values of S_1 correspond to small values of $S_2 = s - S_1$. The conditional distribution of S_1 given $S = s$ is hypergeometric$(n_1 + n_2, n_1, s)$. Thus the conditional p-value is

$$p(s_1, s_2) = \sum_{j=s_1}^{\min\{n_1,s\}} f(j|s),$$

the sum of hypergeometric probabilities. The test defined by this p-value is called Fisher's Exact Test.