Chapter 8

Hypothesis Testing

"It is a mistake to confound strangeness with mystery."

Sherlock Holmes
A Study in Scarlet

8.1 Introduction

In Chapter 7 we studied a method of inference called point estimation. Now we move to another inference method, hypothesis testing. Reflecting the need both to find and to evaluate hypothesis tests, this chapter is divided into two parts, as was Chapter 7. We begin with the definition of a statistical hypothesis.

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

The definition of a hypothesis is rather general, but the important point is that a hypothesis makes a statement about the population. 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.

If θ denotes a population parameter, the general format of the null and alternative hypotheses is $H_0 \colon \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$, where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. For example, if θ denotes the average change in a patient's blood pressure after taking a drug, an experimenter might be interested in testing $H_0 \colon \theta = 0$ versus $H_1 : \theta \neq 0$. The null hypothesis states that, on the average, the drug has no effect on blood pressure, and the alternative hypothesis states that there is some effect. This common situation, in which H_0 states that a treatment has no effect, has led to the term "null" hypothesis. As another example, a consumer might be interested in the proportion of defective items produced by a supplier. If θ denotes the proportion of defective items, the consumer might wish to test $H_0 \colon \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$. The value θ_0 is the maximum acceptable proportion of defective items, and H_0 states that the proportion of defectives is unacceptably high. Problems in which the hypotheses concern the quality of a product are called acceptance sampling problems.

In a hypothesis testing problem, after observing the sample the experimenter must decide either to accept H_0 as true or to reject H_0 as false and decide H_1 is true.

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.*

On a philosophical level, some people worry about the distinction between "rejecting H_0" and "accepting H_1." In the first case, there is nothing implied about what state the experimenter *is* accepting, only that the state defined by H_0 is being rejected. Similarly, a distinction can be made between "accepting H_0" and "not rejecting H_0." The first phrase implies that the experimenter is willing to assert the state of nature specified by H_0, while the second phrase implies that the experimenter really does not believe H_0 but does not have the evidence to reject it. For the most part, we will not be concerned with these issues. We view a hypothesis testing problem as a problem in which one of two actions is going to be taken—the actions being the assertion of H_0 or H_1.

Typically, a hypothesis test is specified in terms of a *test statistic* $W(X_1, \ldots, X_n) = W(\mathbf{X})$, a function of the sample. For example, a test might specify that H_0 is to be rejected if $\bar{X}$, the sample mean, is greater than 3. In this case $W(\mathbf{X}) = \bar{X}$ is the test statistic and the rejection region is $\{(x_1, \ldots, x_n) : \bar{x} > 3\}$. In Section 8.2, methods of choosing test statistics and rejection regions are discussed. Criteria for evaluating tests are introduced in Section 8.3. As with point estimators, the methods of finding tests carry no guarantees; the tests they yield must be evaluated before their worth is established.

8.2 Methods of Finding Tests

We will detail four methods of finding test procedures, procedures that are useful in different situations and take advantage of different aspects of a problem. We start with a very general method, one that is almost always applicable and is also optimal in some cases.

8.2.1 Likelihood Ratio Tests

The likelihood ratio method of hypothesis testing is related to the maximum likelihood estimators discussed in Section 7.2.2, and likelihood ratio tests are as widely applicable as maximum likelihood estimation. Recall that if $X_1, \ldots, X_n$ is a random sample from a population with pdf or pmf $f(x|\theta)$ (θ may be a vector), the likelihood function is defined as

$$L(\theta|x_1,\ldots,x_n) = L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Let Θ denote the entire parameter space. Likelihood ratio tests are defined as follows.

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

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

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

The rationale behind LRTs may best be understood in the situation in which $f(x|\theta)$ is the pmf of a discrete random variable. In this case, the numerator of $\lambda(\mathbf{x})$ is the maximum probability of the observed sample, the maximum being computed over parameters in the null hypothesis. (See Exercise 8.4.) The denominator of $\lambda(\mathbf{x})$ 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. Methods for selecting the number c are discussed in Section 8.3.

If we think of doing the maximization over both the entire parameter space (unrestricted maximization) and a subset of the parameter space (restricted maximization), then the correspondence between LRTs and MLEs becomes more clear. Suppose $\hat{\theta}$, an MLE of θ, exists; $\hat{\theta}$ is obtained by doing an unrestricted maximization of $L(\theta|\mathbf{x})$. 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. That is, $\hat{\theta}_0 = \hat{\theta}_0(\mathbf{x})$ is the value of $\theta \in \Theta_0$ that maximizes $L(\theta|\mathbf{x})$. Then, the LRT statistic is

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

Example 8.2.2 (Normal LRT) Let $X_1,\ldots,X_n$ be a random sample from a $\mathrm{n}(\theta,1)$ population. Consider testing $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$. Here θ_0 is a number fixed by the experimenter prior to the experiment. Since there is only one value of θ specified by H_0, the numerator of $\lambda(\mathbf{x})$ is $L(\theta_0|\mathbf{x})$. In Example 7.2.5 the (unrestricted) MLE of θ was found to be $\bar{X}$, the sample mean. Thus the denominator of $\lambda(\mathbf{x})$ is $L(\bar{x}|\mathbf{x})$. So the LRT statistic is

$$(8.2.1) \qquad \lambda(\mathbf{x}) = \frac{(2\pi)^{-n/2}\exp\left[-\sum_{i=1}^{n}(x_i-\theta_0)^2/2\right]}{(2\pi)^{-n/2}\exp\left[-\sum_{i=1}^{n}(x_i-\bar{x})^2/2\right]}$$

$$= \exp\left[\left(-\sum_{i=1}^{n}(x_i-\theta_0)^2+\sum_{i=1}^{n}(x_i-\bar{x})^2\right)/2\right].$$

The expression for $\lambda(\mathbf{x})$ can be simplified by noting that

$$\sum_{i=1}^{n}(x_i-\theta_0)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\theta_0)^2.$$

Thus the LRT statistic is

$$(8.2.2) \qquad \lambda(\mathbf{x}) = \exp\left[-n(\bar{x}-\theta_0)^2/2\right].$$

An LRT is a test that rejects H_0 for small values of $\lambda(\mathbf{x})$. From (8.2.2), the rejection region, $\{\mathbf{x}\colon \lambda(\mathbf{x}) \le c\}$, can be written as

$$\{\mathbf{x}\colon |\bar{x}-\theta_0| \ge \sqrt{-2(\log c)/n}\}.$$

As c ranges between 0 and 1, $\sqrt{-2(\log c)/n}$ ranges between 0 and ∞. 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. ||

The analysis in Example 8.2.2 is typical in that first the expression for $\lambda(\mathbf{X})$ from Definition 8.2.1 is found, as we did in (8.2.1). Then the description of the rejection region is simplified, if possible, to an expression involving a simpler statistic, $|\bar{X}-\theta_0|$ in the example.

Example 8.2.3 (Exponential LRT) Let $X_1,\ldots,X_n$ be a random sample from an exponential population with pdf

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & x \ge \theta \\ 0 & x < \theta, \end{cases}$$

where $-\infty < \theta < \infty$. The likelihood function is

$$L(\theta|\mathbf{x}) = \begin{cases} e^{-\Sigma x_i + n\theta} & \theta \le x_{(1)} \\ 0 & \theta > x_{(1)}. \end{cases} \qquad (x_{(1)} = \min x_i)$$

Consider testing $H_0\colon \theta \le \theta_0$ versus $H_1\colon \theta > \theta_0$, where θ_0 is a value specified by the experimenter. Clearly $L(\theta|\mathbf{x})$ is an increasing function of θ on $-\infty < \theta \le x_{(1)}$. Thus, the denominator of $\lambda(\mathbf{x})$, the unrestricted maximum of $L(\theta|\mathbf{x})$, is

$$L(x_{(1)}|\mathbf{x}) = e^{-\Sigma x_i + n x_{(1)}}.$$

If $x_{(1)} \le \theta_0$, the numerator of $\lambda(\mathbf{x})$ is also $L(x_{(1)}|\mathbf{x})$. But since we are maximizing $L(\theta|\mathbf{x})$ over $\theta \le \theta_0$, the numerator of $\lambda(\mathbf{x})$ is $L(\theta_0|\mathbf{x})$ if $x_{(1)} > \theta_0$. Therefore, the likelihood ratio test statistic is

$$\lambda(\mathbf{x}) = \begin{cases} 1 & x_{(1)} \le \theta_0 \\ e^{-n(x_{(1)}-\theta_0)} & x_{(1)} > \theta_0. \end{cases}$$

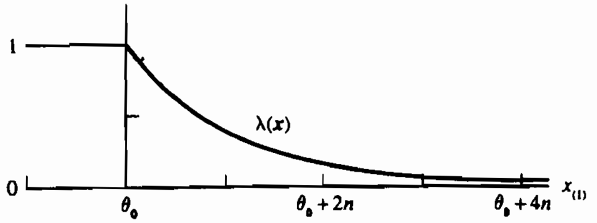

Figure 8.2.1. $\lambda(\mathbf{x})$, *a function only of* $x_{(1)}$.

A graph of $\lambda(\mathbf{x})$ is shown in Figure 8.2.1. An LRT, a test that rejects H_0 if $\lambda(\mathbf{X}) \leq c$, is a test with rejection region $\{\mathbf{x}\colon x_{(1)} \geq \theta_0 - \frac{\log c}{n}\}$. Note that the rejection region depends on the sample only through the sufficient statistic $X_{(1)}$. That this is generally the case will be seen in Theorem 8.2.4. ||

Example 8.2.3 again illustrates the point, expressed in Section 7.2.2, that differentiation of the likelihood function is not the only method of finding an MLE. In Example 8.2.3, $L(\theta|\mathbf{x})$ is not differentiable at $\theta = x_{(1)}$.

If $T(\mathbf{X})$ is a sufficient statistic for θ with pdf or pmf $g(t|\theta)$, then we might consider constructing an LRT based on T and its likelihood function $L^*(\theta|t) = g(t|\theta)$, rather than on the sample $\mathbf{X}$ and its likelihood function $L(\theta|\mathbf{x})$. Let $\lambda^*(t)$ denote the likelihood ratio test statistic based on T. Given the intuitive notion that all the information about θ in $\mathbf{x}$ is contained in $T(\mathbf{x})$, the test based on T should be as good as the test based on the complete sample $\mathbf{X}$. In fact the tests are equivalent.

Theorem 8.2.4 *If $T(\mathbf{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(\mathbf{x})$ are the LRT statistics based on T and $\mathbf{X}$, respectively, then $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$ for every $\mathbf{x}$ in the sample space.*

Proof: From the Factorization Theorem (Theorem 6.2.6), the pdf or pmf of $\mathbf{X}$ can be written as $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$, where $g(t|\theta)$ is the pdf or pmf of T and $h(\mathbf{x})$ does not depend on θ. Thus

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{\sup\limits_{\Theta_0} L(\theta|\mathbf{x})}{\sup\limits_{\Theta} L(\theta|\mathbf{x})} \\ &= \frac{\sup\limits_{\Theta_0} f(\mathbf{x}|\theta)}{\sup\limits_{\Theta} f(\mathbf{x}|\theta)} \\ &= \frac{\sup\limits_{\Theta_0} g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sup\limits_{\Theta} g(T(\mathbf{x})|\theta)h(\mathbf{x})} \qquad (T \text{ is sufficient})\end{aligned}$$

$$= \frac{\sup_{\Theta_0} g(T(\mathbf{x})|\theta)}{\sup_{\Theta} g(T(\mathbf{x})|\theta)} \qquad (h \text{ does not depend on } \theta)$$

$$= \frac{\sup_{\Theta_0} L^*(\theta|T(\mathbf{x}))}{\sup_{\Theta} L^*(\theta|T(\mathbf{x}))} \qquad (g \text{ is the pdf or pmf of } T)$$

$$= \lambda^*(T(\mathbf{x})). \qquad \square$$

The comment after Example 8.2.2 was that, after finding an expression for $\lambda(\mathbf{x})$, we try to simplify that expression. In light of Theorem 8.2.4, one interpretation of this comment is that the simplified expression for $\lambda(\mathbf{x})$ should depend on $\mathbf{x}$ only through $T(\mathbf{x})$ if $T(\mathbf{X})$ is a sufficient statistic for θ.

Example 8.2.5 (LRT and sufficiency) In Example 8.2.2, we can recognize that $\bar{X}$ is a sufficient statistic for θ. We could use the likelihood function associated with $\bar{X}$ ($\bar{X} \sim \mathrm{n}(\theta, \frac{1}{n})$) to more easily reach the conclusion that a likelihood ratio test of $H_0\colon \theta = \theta_0$ versus $H_1\colon \theta \neq \theta_0$ rejects H_0 for large values of $|\bar{X} - \theta_0|$.

Similarly in Example 8.2.3, $X_{(1)} = \min X_i$ is a sufficient statistic for θ. The likelihood function of $X_{(1)}$ (the pdf of $X_{(1)}$) is

$$L^*(\theta|x_{(1)}) = \begin{cases} ne^{-n(x_{(1)}-\theta)} & \theta \leq x_{(1)} \\ 0 & \theta > x_{(1)}. \end{cases}$$

This likelihood could also be used to derive the fact that a likelihood ratio test of $H_0\colon \theta \leq \theta_0$ versus $H_1\colon \theta > \theta_0$ rejects H_0 for large values of $X_{(1)}$. ||

Likelihood ratio tests are also useful in situations where there are *nuisance* parameters, that is, parameters that are present in a model but are not of direct inferential interest. The presence of such nuisance parameters does not affect the LRT construction method but, as might be expected, the presence of nuisance parameters might lead to a different test.

Example 8.2.6 (Normal LRT with unknown variance) Suppose $X_1, \ldots, X_n$ are a random sample from a $\mathrm{n}(\mu, \sigma^2)$, and an experimenter is interested only in inferences about μ, such as testing $H_0\colon \mu \leq \mu_0$ versus $H_1\colon \mu > \mu_0$. Then the parameter σ^2 is a nuisance parameter. The LRT statistic is

$$\lambda(\mathbf{x}) = \frac{\max_{\{\mu,\sigma^2:\mu\leq\mu_0,\sigma^2\geq 0\}} L(\mu,\sigma^2|\mathbf{x})}{\max_{\{\mu,\sigma^2:-\infty<\mu<\infty,\sigma^2\geq 0\}} L(\mu,\sigma^2|\mathbf{x})}$$

$$= \frac{\max_{\{\mu,\sigma^2:\mu\leq\mu_0,\sigma^2\geq 0\}} L(\mu,\sigma^2|\mathbf{x})}{L(\hat{\mu},\hat{\sigma}^2|\mathbf{x})},$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the MLEs of μ and σ^2 (see Example 7.2.11). Furthermore, if $\hat{\mu} \leq \mu_0$, then the restricted maximum is the same as the unrestricted maximum, while if $\hat{\mu} > \mu_0$, the restricted maximum is $L(\mu_0, \hat{\sigma}_0^2|\mathbf{x})$, where $\hat{\sigma}_0^2 = \Sigma\,(x_i - \mu_0)^2/n$. Thus

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{\mu} \le \mu_0 \\ \frac{L(\mu_0, \hat{\sigma}_0^2|\mathbf{x})}{L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x})} & \text{if } \hat{\mu} > \mu_0. \end{cases}$$

With some algebra, it can be shown that the test based on $\lambda(\mathbf{x})$ is equivalent to a test based on Student's t statistic. Details are left to Exercise 8.37. (Exercises 8.38–8.42 also deal with nuisance parameter problems.) ||

8.2.2 Bayesian Tests

Hypothesis testing problems may also be formulated in a Bayesian model. Recall from Section 7.2.3 that a Bayesian model includes not only the sampling distribution $f(\mathbf{x}|\theta)$ but also the prior distribution $\pi(\theta)$, with the prior distribution reflecting the experimenter's opinion about the parameter θ prior to sampling.

The Bayesian paradigm prescribes that the sample information be combined with the prior information using Bayes' Theorem to obtain the posterior distribution $\pi(\theta|\mathbf{x})$. All inferences about θ are now based on the posterior distribution.

In a hypothesis testing problem, the posterior distribution may be used to calculate the probabilities that H_0 and H_1 are true. Remember, $\pi(\theta|\mathbf{x})$ is a probability distribution for a random variable. Hence, the posterior probabilities $P(\theta \in \Theta_0|\mathbf{x}) = P(H_0$ is true$|\mathbf{x})$ and $P(\theta \in \Theta_0^c|\mathbf{x}) = P(H_1$ is true$|\mathbf{x})$ may be computed.

The probabilities $P(H_0$ is true$|\mathbf{x})$ and $P(H_1$ is true$|\mathbf{x})$ are not meaningful to the classical statistician. The classical statistician considers θ to be a fixed number. Consequently, a hypothesis is *either true or false.* If $\theta \in \Theta_0$, $P(H_0$ is true$|\mathbf{x}) = 1$ and $P(H_1$ is true$|\mathbf{x}) = 0$ for all values of $\mathbf{x}$. If $\theta \in \Theta_0^c$, these values are reversed. Since these probabilities are unknown (since θ is unknown) and do not depend on the sample $\mathbf{x}$, they are not used by the classical statistician. In a Bayesian formulation of a hypothesis testing problem, these probabilities depend on the sample $\mathbf{x}$ and can give useful information about the veracity of H_0 and H_1.

One way a Bayesian hypothesis tester may choose to use the posterior distribution is to decide to accept H_0 as true if $P(\theta \in \Theta_0|\mathbf{X}) \ge P(\theta \in \Theta_0^c|\mathbf{X})$ and to reject H_0 otherwise. In the terminology of the previous sections, the test statistic, a function of the sample, is $P(\theta \in \Theta_0^c|\mathbf{X})$ and the rejection region is $\{\mathbf{x}\colon P(\theta \in \Theta_0^c|\mathbf{x}) > \frac{1}{2}\}$. Alternatively, if the Bayesian hypothesis tester wishes to guard against falsely rejecting H_0, he may decide to reject H_0 only if $P(\theta \in \Theta_0^c|\mathbf{X})$ is greater than some large number, .99 for example.

Example 8.2.7 (Normal Bayesian test) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$ and let the prior distribution on θ be $\mathrm{n}(\mu, \tau^2)$, where σ^2, μ, and τ^2 are known. Consider testing $H_0\colon \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$. From Example 7.2.16, the posterior $\pi(\theta|\bar{x})$ 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 $P(\theta \in \Theta_0|\mathbf{X}) \ge P(\theta \in \Theta_0^c|\mathbf{X})$, then we will accept H_0 if and only if

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

Since $\pi(\theta|\mathbf{x})$ is symmetric, this is true if and only if the mean of $\pi(\theta|\mathbf{x})$ is less than or equal to θ_0. 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. In particular, if $\mu = \theta_0$ so that prior to experimentation probability $\frac{1}{2}$ is assigned to both H_0 and H_1, then H_0 will be accepted as true if $\bar{x} \leq \theta_0$ and H_1 accepted otherwise. ||

Other methods that use the posterior distribution to make inferences in hypothesis testing problems are discussed in Section 8.3.5.

8.2.3 Union–Intersection and Intersection–Union Tests

In some situations, tests for complicated null hypotheses can be developed from tests for simpler null hypotheses. We discuss two related methods.

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

$$H_0\colon \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma. \tag{8.2.3}$$

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 $H_{0\gamma}\colon \theta \in \Theta_\gamma$ versus $H_{1\gamma}\colon \theta \in \Theta_\gamma^c$. Say the rejection region for the test of $H_{0\gamma}$ is $\{\mathbf{x}\colon T_\gamma(\mathbf{x}) \in R_\gamma\}$. Then the rejection region for the union–intersection test is

$$\bigcup_{\gamma \in \Gamma} \{\mathbf{x}:\ T_\gamma(\mathbf{x}) \in R_\gamma\}. \tag{8.2.4}$$

The rationale is simple. If any one of the hypotheses $H_{0\gamma}$ is rejected, then H_0, which, by (8.2.3), is true only if $H_{0\gamma}$ is true for every γ, must also be rejected. Only if each of the hypotheses $H_{0\gamma}$ is accepted as true will the intersection H_0 be accepted as true.

In some situations a simple expression for the rejection region of a union–intersection test can be found. In particular, suppose that each of the individual tests has a rejection region of the form $\{\mathbf{x}\colon T_\gamma(\mathbf{x}) > c\}$, where c does not depend on γ. The rejection region for the union–intersection test, given in (8.2.4), can be expressed as

$$\bigcup_{\gamma \in \Gamma} \{\mathbf{x}\colon T_\gamma(\mathbf{x}) > c\} = \{\mathbf{x}\colon \sup_{\gamma \in \Gamma} T_\gamma(\mathbf{x}) > c\}.$$

Thus the test statistic for testing H_0 is $T(\mathbf{x}) = \sup_{\gamma \in \Gamma} T_\gamma(\mathbf{x})$. Some examples in which $T(\mathbf{x})$ has a simple formula may be found in Chapter 11.

Example 8.2.8 (Normal union–intersection test) Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population. Consider testing $H_0\colon \mu = \mu_0$ versus $H_1\colon \mu \neq \mu_0$, where μ_0 is a specified number. We can write H_0 as the intersection of two sets,

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

The LRT of $H_{0\text{L}}: \mu \le \mu_0$ versus $H_{1\text{L}}: \mu > \mu_0$ is

$$\text{reject } H_{0\text{L}}: \mu \le \mu_0 \text{ in favor of } H_{1\text{L}}: \mu > \mu_0 \text{ if } \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \ge t_{\text{L}}$$

(see Exercise 8.37). Similarly, the LRT of $H_{0\text{U}}: \mu \ge \mu_0$ versus $H_{1\text{U}}: \mu < \mu_0$ is

$$\text{reject } H_{0\text{U}}: \mu \ge \mu_0 \text{ in favor of } H_{1\text{U}}: \mu < \mu_0 \text{ if } \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \le t_{\text{U}}.$$

Thus the union–intersection test of $H_0: \mu = \mu_0$ versus $H_1: \mu \ne \mu_0$ formed from these two LRTs is

$$\text{reject } H_0 \text{ if } \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \ge t_{\text{L}} \text{ or } \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \le t_{\text{U}}.$$

If $t_{\text{L}} = -t_{\text{U}} \ge 0$, the union–intersection test can be more simply expressed as

$$\text{reject } H_0 \text{ if } \frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} \ge t_{\text{L}}.$$

It turns out that this union–intersection test is also the LRT for this problem (see Exercise 8.38) and is called the two-sided t test. ||

The union–intersection method of test construction is useful if the null hypothesis is conveniently expressed as an intersection. Another method, the *intersection–union method*, may be useful if the null hypothesis is conveniently expressed as a union. Suppose we wish to test the null hypothesis

$$H_0: \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma. \tag{8.2.5}$$

Suppose that for each $\gamma \in \Gamma, \{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\}$ 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 intersection–union test of H_0 versus H_1 is

$$\bigcap_{\gamma \in \Gamma} \{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\}. \tag{8.2.6}$$

From (8.2.5), H_0 is false if and only if *all* of the $H_{0\gamma}$ are false, so H_0 can be rejected if and only if each of the individual hypotheses $H_{0\gamma}$ can be rejected. Again, the test can be greatly simplified if the rejection regions for the individual hypotheses are all of the form $\{\mathbf{x}: T_\gamma(\mathbf{x}) \ge c\}$ (c independent of γ). In such cases, the rejection region for H_0 is

$$\bigcap_{\gamma \in \Gamma} \{\mathbf{x}: T_\gamma(\mathbf{x}) \ge c\} = \{\mathbf{x}: \inf_{\gamma \in \Gamma} T_\gamma(\mathbf{x}) \ge c\}.$$

Here, the intersection-union test statistic is $\inf_{\gamma \in \Gamma} T_\gamma(\mathbf{X})$, and the test rejects H_0 for large values of this statistic.

Example 8.2.9 (Acceptance sampling) The topic of acceptance sampling provides an extremely useful application of an intersection–union test, as this example will illustrate. (See Berger 1982 for a more detailed treatment of this problem.)

Two parameters that are important in assessing the quality of 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 .95, and the fabric is acceptable only if it meets both of these standards. This can be modeled with the hypothesis test

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

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

Suppose $X_1, \ldots, X_n$ are measurements of breaking strength for n samples and are assumed to be iid $\mathrm{n}(\theta_1, \sigma^2)$. 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 tests, denoted by $Y_1, \ldots, Y_m$, where $Y_i = 1$ if the ith 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 $H_{02} \colon \theta_2 \leq .95$ if $\sum_{i=1}^{m} Y_i > b$ (see Exercise 8.3). Putting all of this together, the rejection region for the intersection–union test is given by

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

Thus the intersection–union test decides the product is acceptable, that is, H_1 is true, if and only if it decides that each of the individual parameters meets its standard, that is, H_{1i} is true. If more than two parameters define a product's quality, individual tests for each parameter can be combined, by means of the intersection–union method, to yield an overall test of the product's quality. ||

8.3 Methods of Evaluating Tests

In deciding to accept or reject the null hypothesis H_0, an experimenter might be making a mistake. Usually, hypothesis tests are evaluated and compared through their probabilities of making mistakes. In this section we discuss how these error probabilities can be controlled. In some cases, it can even be determined which tests have the smallest possible error probabilities.

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. These two types of errors traditionally have been given the non-mnemonic names, 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 8.3.1.

Suppose R denotes the rejection region for a test. Then for $\theta \in \Theta_0$, the test will make a mistake if $\mathbf{x} \in R$, so the probability of a Type I Error is $P_\theta(\mathbf{X} \in R)$. For

Table 8.3.1. *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

$\theta \in \Theta_0^c$, the probability of a Type II Error is $P_\theta(\mathbf{X} \in R^c)$. This switching from R to R^c is a bit confusing, but, if we realize that $P_\theta(\mathbf{X} \in R^c) = 1 - P_\theta(\mathbf{X} \in R)$, then the function of θ, $P_\theta(\mathbf{X} \in R)$, contains all the information about the test with rejection region R. We have

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

This consideration leads to the following definition.

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

The ideal power function is 0 for all $\theta \in \Theta_0$ and 1 for all $\theta \in \Theta_0^c$. Except in trivial situations, this ideal cannot be attained. 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 8.3.2 (Binomial power function) Let $X \sim$ binomial$(5, \theta)$. Consider testing $H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$. Consider first the test that rejects H_0 if and only if all "successes" are observed. The power function for this test is

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

The graph of $\beta_1(\theta)$ is in Figure 8.3.1. In examining this power function, we might decide that although the probability of a Type I Error is acceptably low ($\beta_1(\theta) \leq (\frac{1}{2})^5 = .0312$) for all $\theta \leq \frac{1}{2}$, the probability of a Type II Error is too high ($\beta_1(\theta)$ is too small) for most $\theta > \frac{1}{2}$. The probability of a Type II Error is less than $\frac{1}{2}$ only if $\theta > (\frac{1}{2})^{1/5} = .87$. 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

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

The graph of $\beta_2(\theta)$ is also in Figure 8.3.1. It can be seen in Figure 8.3.1 that the second test has achieved a smaller Type II Error probability in that $\beta_2(\theta)$ is larger for $\theta > \frac{1}{2}$. But the Type I Error probability is larger for the second test; $\beta_2(\theta)$ is larger for $\theta \leq \frac{1}{2}$. If a choice is to be made between these two tests, the researcher must decide which error structure, that described by $\beta_1(\theta)$ or that described by $\beta_2(\theta)$, is more acceptable. ||

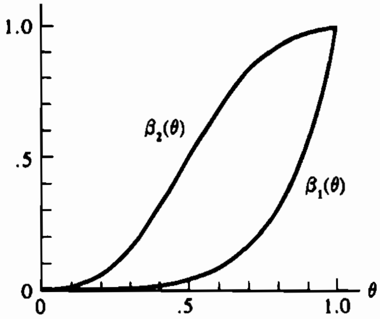

Figure 8.3.1. *Power functions for Example 8.3.2*

Example 8.3.3 (Normal power function) Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, σ^2 known. An LRT of $H_0\colon \theta \le \theta_0$ versus $H_1\colon \theta > \theta_0$ is a test that rejects H_0 if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > c$ (see Exercise 8.37). The constant c can be any positive number. The power function of this test is

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

where Z is a standard normal random variable, since $(\bar{X} - \theta)/(\sigma/\sqrt{n}) \sim \mathrm{n}(0,1)$. As θ increases from $-\infty$ to ∞, it is easy to see that this normal probability increases from 0 to 1. Therefore, it follows that $\beta(\theta)$ is an increasing function of θ, with

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

A graph of $\beta(\theta)$ for $c = 1.28$ is given in Figure 8.3.2. ||

Typically, the power function of a test will depend on the sample size n. If n can be chosen by the experimenter, consideration of the power function might help determine what sample size is appropriate in an experiment.

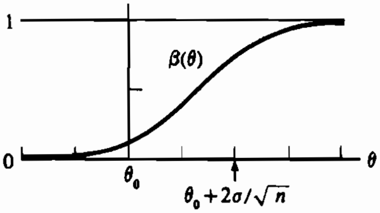

Figure 8.3.2. *Power function for Example 8.3.3*

Example 8.3.4 (Continuation of Example 8.3.3) Suppose the experimenter wishes to have a maximum Type I Error probability of .1. Suppose, in addition, the experimenter wishes to have a maximum Type II Error probability of .2 if $\theta \geq \theta_0 + \sigma$. We now show how to choose c and n to achieve these goals, using a test that rejects $H_0 : \theta \leq \theta_0$ if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > c$. As noted above, the power function of such a test is

$$\beta(\theta) = P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right).$$

Because $\beta(\theta)$ is increasing in θ, the requirements will be met if

$$\beta(\theta_0) = .1 \quad \text{and} \quad \beta(\theta_0 + \sigma) = .8.$$

By choosing $c = 1.28$, we achieve $\beta(\theta_0) = P(Z > 1.28) = .1$, regardless of n. Now we wish to choose n so that $\beta(\theta_0+\sigma) = P(Z > 1.28 - \sqrt{n}) = .8$. But, $P(Z > -.84) = .8$. So setting $1.28 - \sqrt{n} = -.84$ and solving for n yield $n = 4.49$. Of course n must be an integer. 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. The following two terms are useful when discussing tests that control Type I Error probabilities.

Definition 8.3.5 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 8.3.6 For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a *level α test* if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

Some authors do not make the distinction between the terms *size* and *level* that we have made, and sometimes these terms are used interchangeably. But according to our definitions, the set of level α tests contains the set of size α tests. Moreover, the distinction becomes important in complicated models and complicated testing situations, where it is often computationally impossible to construct a size α test. In such situations, an experimenter must be satisfied with a level α test, realizing that some compromises may be made. We will see some examples, especially in conjunction with union–intersection and intersection–union tests.

Experimenters commonly specify the level of the test they wish to use, with typical choices being $\alpha = .01$, .05, and .10. 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. If this approach is taken, the experimenter should specify the null and alternative hypotheses so that it is most important to control the Type I Error probability. For example, suppose an experimenter expects an experiment to give support to a particular hypothesis, but she does not wish to make the assertion unless the data really do give convincing support. The test can be set up so that the alternative hypothesis is the one that she expects the data to support, and hopes to prove. (The alternative hypothesis is sometimes called the *research hypothesis* in this context.) By using a level α test with small α, the experimenter is guarding against saying the data support the research hypothesis when it is false.

The methods of Section 8.2 usually yield test statistics and general forms for rejection regions. However, they do not generally lead to one specific test. For example, an LRT (Definition 8.2.1) is one that rejects H_0 if $\lambda(\mathbf{X}) \le c$, but c was unspecified, so not one but an entire class of LRTs is defined, one for each value of c. The restriction to size α tests may now lead to the choice of one out of the class of tests.

Example 8.3.7 (Size of LRT) In general, a size α LRT is constructed by choosing c such that $\sup_{\theta\in\Theta_0} P_\theta(\lambda(\mathbf{X}) \le c) = \alpha$. How that c is determined depends on the particular problem. For example, in Example 8.2.2, Θ_0 consists of the single point $\theta = \theta_0$ and $\sqrt{n}(\bar{X} - \theta_0) \sim \mathrm{n}(0,1)$ if $\theta = \theta_0$. So the test

$$\text{reject } H_0 \text{ if } \left|\bar{X} - \theta_0\right| \ge z_{\alpha/2}/\sqrt{n},$$

where $z_{\alpha/2}$ satisfies $P(Z \ge z_{\alpha/2}) = \alpha/2$ with $Z \sim \mathrm{n}(0,1)$, is the size α LRT. Specifically, this corresponds to choosing $c = \exp(-z^2_{\alpha/2}/2)$, but this is not an important point.

For the problem described in Example 8.2.3, finding a size α LRT is complicated by the fact that the null hypothesis $H_0\colon \theta \le \theta_0$ consists of more than one point. The LRT rejects H_0 if $X_{(1)} \ge c$, where c is chosen so that this is a size α test. But if $c = (-\log\alpha)/n + \theta_0$, then

$$P_{\theta_0}\left(X_{(1)} \ge c\right) = e^{-n(c-\theta_0)} = \alpha.$$

Since θ is a location parameter for $X_{(1)}$,

$$P_\theta\left(X_{(1)} \ge c\right) \le P_{\theta_0}\left(X_{(1)} \ge c\right) \quad \text{for any } \theta \le \theta_0.$$

Thus

$$\sup_{\theta\in\Theta_0} \beta(\theta) = \sup_{\theta\le\theta_0} P_\theta(X_{(1)} \ge c) = P_{\theta_0}(X_{(1)} \ge c) = \alpha$$

and this c yields the size α LRT. ||

A note on notation: In the above example we used the notation $z_{\alpha/2}$ to denote the point having probability $\alpha/2$ to the right of it for a standard normal pdf. We will use this notation in general, not just for the normal but for other distributions as well (defining what we need to for clarity's sake). For example, the point z_α satisfies $P(Z > z_\alpha) = \alpha$, where $Z \sim \mathrm{n}(0,1)$; $t_{n-1,\alpha/2}$ satisfies $P(T_{n-1} > t_{n-1,\alpha/2}) = \alpha/2$, where $T_{n-1} \sim t_{n-1}$; and $\chi^2_{p,1-\alpha}$ satisfies $P(\chi^2_p > \chi^2_{p,1-\alpha}) = 1 - \alpha$, where χ^2_p is a chi squared random variable with p degrees of freedom. Points like $z_{\alpha/2}, z_\alpha, t_{n-1,\alpha/2}$, and $\chi^2_{p,1-\alpha}$ are known as *cutoff points.*

Example 8.3.8 (Size of union–intersection test) The problem of finding a size α union–intersection test in Example 8.2.8 involves finding constants t_{L} and t_{U} such that

$$\sup_{\theta \in \Theta_0} P_\theta \left(\frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \geq t_{\mathrm{L}} \quad \text{or} \quad \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \leq t_{\mathrm{U}} \right) = \alpha.$$

But for any $(\mu, \sigma^2) = \theta \in \Theta_0$, $\mu = \mu_0$ and thus $(\bar{X} - \mu_0)/\sqrt{S^2/n}$ has a Student's t distribution with $n-1$ degrees of freedom. So any choice of $t_{\mathrm{U}} = t_{n-1,1-\alpha_1}$ and $t_{\mathrm{L}} = t_{n-1,\alpha_2}$, 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 $t_{\mathrm{L}} = -t_{\mathrm{U}} = t_{n-1,\alpha/2}$. ||

Other than α levels, 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 $\theta \in \Theta_0$. All of the power functions in Figures 8.3.1 and 8.3.2 have this property, yielding tests that are called unbiased.

Definition 8.3.9 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 8.3.10 (Conclusion of Example 8.3.3) An LRT of $H_0\colon \theta \leq \theta_0$ versus $H_1\colon \theta > \theta_0$ has power function

$$\beta(\theta) = P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} \right),$$

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

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

and, hence, that the test is unbiased. ||

In most problems there are many unbiased tests. (See Exercise 8.45.) Likewise, there are many size α tests, likelihood ratio tests, etc. In some cases we have imposed enough restrictions to narrow consideration to one test. For the two problems in Example 8.3.7, there is only one size α likelihood ratio test. In other cases there remain many tests from which to choose. We discussed only the one that rejects H_0 for large values of T. In the following sections we will discuss other criteria for selecting one out of a class of tests, criteria that are all related to the power functions of the tests.

8.3.2 Most Powerful Tests

In previous sections we have described various classes of hypothesis tests. Some of these classes control the probability of a Type I Error; for example, level α tests have Type I Error probabilities at most α for all $\theta \in \Theta_0$. A good test in such a class would also have a small Type II Error probability, that is, a large power function for $\theta \in \Theta_0^c$. If one test had a smaller Type II Error probability than all other tests in the class, it would certainly be a strong contender for the best test in the class, a notion that is formalized in the next definition.

Definition 8.3.11 Let $\mathcal{C}$ be a class of tests for testing $H_0\colon \theta \in \Theta_0$ versus $H_1\colon \theta \in \Theta_0^c$. 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 Definition 8.3.11 is then called a UMP level α test. For this test to be interesting, restriction to the class $\mathcal{C}$ must involve some restriction on the Type I Error probability. A minimization of the Type II Error probability without some control of the Type I Error probability is not very interesting. (For example, a test that rejects H_0 with probability 1 will never make a Type II Error. See Exercise 8.16.)

The requirements in Definition 8.3.11 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. Thus, we would like to be able to identify UMP tests if they exist. The following famous theorem clearly describes which tests are UMP level α tests in the situation where the null and alternative hypotheses both consist of only one probability distribution for the sample (that is, when both H_0 and H_1 are *simple* hypotheses).

Theorem 8.3.12 (Neyman–Pearson Lemma) *Consider testing $H_0 : \theta = \theta_0$ versus $H_1\colon \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(\mathbf{x}|\theta_i), i = 0, 1$, using a test with rejection region R that satisfies*

$$\mathbf{x} \in R \quad \text{if} \quad f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0)$$

$$\text{(8.3.1)} \qquad \text{and}$$

$$\mathbf{x} \in R^c \quad \text{if} \quad f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0),$$

for some $k \geq 0$, and

$$\text{(8.3.2)} \qquad \alpha = P_{\theta_0}(\mathbf{X} \in R).$$

Then

a. *(Sufficiency) Any test that satisfies (8.3.1) and (8.3.2) is a UMP level α test.*

b. *(Necessity) If there exists a test satisfying (8.3.1) and (8.3.2) with $k > 0$, then every UMP level α test is a size α test (satisfies (8.3.2)) and every UMP level α test satisfies (8.3.1) except perhaps on a set A satisfying $P_{\theta_0}(\mathbf{X} \in A) = P_{\theta_1}(\mathbf{X} \in A) = 0$.*

Proof: We will prove the theorem for the case that $f(\mathbf{x}|\theta_0)$ and $f(\mathbf{x}|\theta_1)$ are pdfs of continuous random variables. The proof for discrete random variables can be accomplished by replacing integrals with sums. (See Exercise 8.21.)

Note first that any test satisfying (8.3.2) is a size α and, hence, a level α test because $\sup_{\theta\in\Theta_0} P_\theta(\mathbf{X} \in R) = P_{\theta_0}(\mathbf{X} \in R) = \alpha$, 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 $\mathbf{x} \in R$ and 0 if $\mathbf{x} \in R^c$. That is, it is the indicator function of the rejection region. Let $\phi(\mathbf{x})$ be the test function of a test satisfying (8.3.1) and (8.3.2). Let $\phi'(\mathbf{x})$ 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 $0 \le \phi'(\mathbf{x}) \le 1$, (8.3.1) implies that $(\phi(\mathbf{x}) - \phi'(\mathbf{x}))(f(\mathbf{x}|\theta_1) - kf(\mathbf{x}|\theta_0)) \ge 0$ for every $\mathbf{x}$ (since $\phi = 1$ if $f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0)$ and $\phi = 0$ if $f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0)$). Thus

$$(8.3.3) \qquad \begin{aligned} 0 &\le \int [\phi(\mathbf{x}) - \phi'(\mathbf{x})][f(\mathbf{x}|\theta_1) - kf(\mathbf{x}|\theta_0)]\, d\mathbf{x} \\ &= \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0)). \end{aligned}$$

Statement (a) is proved by noting that, since ϕ' is a level α test and ϕ is a size α test, $\beta(\theta_0) - \beta'(\theta_0) = \alpha - \beta'(\theta_0) \ge 0$. Thus (8.3.3) and $k \ge 0$ imply that

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

showing that $\beta(\theta_1) \ge \beta'(\theta_1)$ 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 (8.3.1) and (8.3.2), is also a UMP level α test, thus $\beta(\theta_1) = \beta'(\theta_1)$. This fact, (8.3.3), and $k > 0$ imply

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

Now, since ϕ' is a level α test, $\beta'(\theta_0) \le \alpha$. Thus $\beta'(\theta_0) = \alpha$, that is, ϕ' is a size α test, and this also implies that (8.3.3) is an equality in this case. But the nonnegative integrand $(\phi(\mathbf{x}) - \phi'(\mathbf{x}))(f(\mathbf{x}|\theta_1) - kf(\mathbf{x}|\theta_0))$ will have a zero integral only if ϕ' satisfies (8.3.1), except perhaps on a set A with $\int_A f(\mathbf{x}|\theta_i)\, d\mathbf{x} = 0$. This implies that the last assertion in statement (b) is true. □

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

Corollary 8.3.13 *Consider the hypothesis problem posed in Theorem 8.3.12. Suppose $T(\mathbf{X})$ 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*

$$(8.3.4) \qquad \begin{aligned} &t \in S \quad \text{if} \quad g(t|\theta_1) > kg(t|\theta_0) \\ &\text{and} \\ &t \in S^c \quad \text{if} \quad g(t|\theta_1) < kg(t|\theta_0), \end{aligned}$$

for some $k \ge 0$, where

$$(8.3.5) \qquad \alpha = P_{\theta_0}(T \in S).$$

Proof: In terms of the original sample $\mathbf{X}$, the test based on T has the rejection region $R = \{\mathbf{x}: T(\mathbf{x}) \in S\}$. By the Factorization Theorem, the pdf or pmf of $\mathbf{X}$ can be written as $f(\mathbf{x}|\theta_i) = g(T(\mathbf{x})|\theta_i)h(\mathbf{x}), i = 0, 1$, for some nonnegative function $h(\mathbf{x})$. Multiplying the inequalities in (8.3.4) by this nonnegative function, we see that R satisfies

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

and

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

Also, by (8.3.5),

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

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

When we derive a test that satisfies the inequalities (8.3.1) or (8.3.4), and hence is a UMP level α test, it is usually easier to rewrite the inequalities as $f(\mathbf{x}|\theta_1)/f(\mathbf{x}|\theta_0) > k$. (We must be careful about dividing by 0.) This method is used in the following examples.

Example 8.3.14 (UMP binomial test) Let $X \sim \text{binomial}(2, \theta)$. We want to test $H_0: \theta = \frac{1}{2}$ versus $H_1: \theta = \frac{3}{4}$. Calculating the ratios of the pmfs gives

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

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

Note that if $k = \frac{3}{4}$, then (8.3.1) 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 $\alpha = \frac{1}{4}$ test as above. If we reject H_0 for $x = 1$, we get the UMP level $\alpha = \frac{3}{4}$ test as above. ||

Example 8.3.14 also shows that for a discrete distribution, the α level at which a test can be done is a function of the particular pmf with which we are dealing. (No such problem arises in the continuous case. Any α level can be attained.)

Example 8.3.15 (UMP normal test) Let $X_1, \ldots, X_n$ be a random sample from a $\text{n}(\theta, \sigma^2)$ 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 (8.3.4), $g(\bar{x}|\theta_1) > kg(\bar{x}|\theta_0)$, is equivalent to

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

The fact that $\theta_1 - \theta_0 < 0$ was used to obtain this inequality. The right-hand side increases from $-\infty$ to ∞ as k increases from 0 to ∞. Thus, by Corollary 8.3.13, 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 $\bar{X} < c = -\sigma z_\alpha/\sqrt{n} + \theta_0$. This choice of c ensures that (8.3.5) is true. ||

Hypotheses, such as H_0 and H_1 in the Neyman–Pearson Lemma, that specify only one possible distribution for the sample $\mathbf{X}$ 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 8.3.11 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 8.3.16 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 $\{t: g(t|\theta_1) > 0 \text{ or } g(t|\theta_2) > 0\}$. Note that $c/0$ is defined as ∞ if $0 < c$.

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 (see Exercise 8.25).

Theorem 8.3.17 (Karlin–Rubin) *Consider testing $H_0 : \theta \leq \theta_0$ 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. Fix $\theta' > \theta_0$ and consider testing $H_0': \theta = \theta_0$ versus $H_1': \theta = \theta'$. Since the family of pdfs or pmfs of T has an MLR, $\beta(\theta)$ is nondecreasing (see Exercise 8.34), so

i. $\sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$, 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 8.3.13, (i) and (ii) imply that $\beta(\theta') \geq \beta^*(\theta')$, where $\beta^*(\theta)$ is the power function for any other level α test of H_0', that is, any test satisfying $\beta(\theta_0) \leq \alpha$. However, any level α test of H_0 satisfies $\beta^*(\theta_0) \leq \sup_{\theta \in \Theta_0} \beta^*(\theta) \leq \alpha$. Thus, $\beta(\theta') \geq \beta^*(\theta')$ 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 8.3.17, the test that rejects $H_0\colon \theta \geq \theta_0$ in favor of $H_1\colon \theta < \theta_0$ if and only if $T < t_0$ is a UMP level $\alpha = P_{\theta_0}(T < t_0)$ test.

Example 8.3.18 (Continuation of Example 8.3.15) Consider testing $H_0'\colon \theta \geq \theta_0$ versus $H_1'\colon \theta < \theta_0$ 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 (see Exercise 8.25), it follows from Theorem 8.3.17 that the test is a UMP level α test in this problem.

As the power function of this test,

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

is a decreasing function of θ (since θ is a location parameter in the distribution of $\bar{X}$), the value of α is given by $\sup_{\theta \geq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$. ||

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. This tactic should be reminiscent of what we did in Chapter 7, when we restricted attention to unbiased point estimators in order to investigate optimality. We illustrate how restricting attention to the subset consisting of unbiased tests can result in finding a best test.

First we consider an example that illustrates a typical situation in which a UMP level α test does not exist.

Example 8.3.19 (Nonexistence of UMP test) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$, σ^2 known. Consider testing $H_0\colon \theta = \theta_0$ versus $H_1\colon \theta \neq \theta_0$. For a specified value of α, a level α test in this problem is any test that satisfies

$$P_{\theta_0}(\text{reject } H_0) \leq \alpha. \tag{8.3.6}$$

Consider an alternative parameter point $\theta_1 < \theta_0$. The analysis in Example 8.3.18 shows that, among all tests that satisfy (8.3.6), 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 $\int_A f(\mathbf{x}|\theta_i)\,d\mathbf{x} = 0$. 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 $\theta_2 > \theta_0$,

$$\begin{aligned}
\beta_2(\theta_2) &= P_{\theta_2}\left(\bar{X} > \frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\right) \\
&= P_{\theta_2}\left(\frac{\bar{X} - \theta_2}{\sigma/\sqrt{n}} > z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}\right) \\
&> P(Z > z_\alpha) && \left(\begin{array}{c} Z \sim \mathrm{n}(0,1), \\ > \text{ since } \theta_0 - \theta_2 < 0 \end{array}\right) \\
&= P(Z < -z_\alpha) \\
&> P_{\theta_2}\left(\frac{\bar{X} - \theta_2}{\sigma/\sqrt{n}} < -z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}\right) && \left(\begin{array}{c} \text{again, } > \text{ since} \\ \theta_0 - \theta_2 < 0 \end{array}\right) \\
&= P_{\theta_2}\left(\bar{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\right) \\
&= \beta_1(\theta_2).
\end{aligned}$$

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. ||

Example 8.3.19 illustrates again the usefulness of the Neyman–Pearson Lemma. Previously, the *sufficiency* part of the lemma was used to construct UMP level α tests, but to show the nonexistence of a UMP level α test, the *necessity* part of the lemma is used.

Example 8.3.20 (Unbiased test) 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. The power function, $\beta_3(\theta)$, of Test 3, which rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ if and only if

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

as well as $\beta_1(\theta)$ and $\beta_2(\theta)$ from Example 8.3.19, is shown in Figure 8.3.3. Test 3 is actually a UMP unbiased level α test; that is, it is UMP in the class of unbiased tests.

Note that although Test 1 and Test 2 have slightly higher powers than Test 3 for some parameter points, Test 3 has much higher power than Test 1 and Test 2 at other parameter points. For example, $\beta_3(\theta_2)$ is near 1, whereas $\beta_1(\theta_2)$ is near 0. If the interest is in rejecting H_0 for both large and small values of θ, Figure 8.3.3 shows that Test 3 is better overall than either Test 1 or Test 2. ||

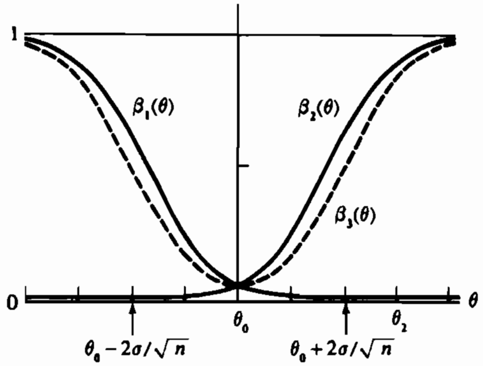

Figure 8.3.3. *Power functions for three tests in Example 8.3.19; $\beta_3(\theta)$ is the power function of an unbiased level $\alpha = .05$ test*

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 of 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.

First consider UITs. Recall that, in this situation, we are testing a null hypothesis of the form $H_0\colon \theta \in \Theta_0$, where $\Theta_0 = \bigcap_{\gamma\in\Gamma} \Theta_\gamma$. To be specific, let $\lambda_\gamma(\mathbf{x})$ be the LRT statistic for testing $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$, and let $\lambda(\mathbf{x})$ be the LRT statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. Then we have the following relationships between the overall LRT and the UIT based on $\lambda_\gamma(\mathbf{x})$.

Theorem 8.3.21 *Consider testing $H_0\colon \theta \in \Theta_0$ versus $H_1\colon \theta \in \Theta_0^c$, where $\Theta_0 = \bigcap_{\gamma\in\Gamma} \Theta_\gamma$ and $\lambda_\gamma(\mathbf{x})$ is defined in the previous paragraph. Define $T(\mathbf{x}) = \inf_{\gamma\in\Gamma}\lambda_\gamma(\mathbf{x})$, and form the UIT with rejection region*

$$\{\mathbf{x}\colon \lambda_\gamma(\mathbf{x}) < c \textit{ for some } \gamma \in \Gamma\} = \{\mathbf{x}\colon T(\mathbf{x}) < c\}.$$

Also consider the usual LRT with rejection region $\{\mathbf{x}\colon \lambda(\mathbf{x}) < c\}$. Then

a. *$T(\mathbf{x}) \geq \lambda(\mathbf{x})$ for every $\mathbf{x}$;*

b. *If $\beta_T(\theta)$ and $\beta_\lambda(\theta)$ are the power functions for the tests based on T and λ, respectively, then $\beta_T(\theta) \leq \beta_\lambda(\theta)$ 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 $\mathbf{x}$,

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

because the region of maximization is bigger for the individual λ_γ. Thus $T(\mathbf{x}) = \inf_{\gamma\in\Gamma}\lambda_\gamma(\mathbf{x}) \geq \lambda(\mathbf{x})$, proving (a). By (a), $\{\mathbf{x}\colon T(\mathbf{x}) < c\} \subset \{\mathbf{x}\colon \lambda(\mathbf{x}) < c\}$, so

$$\beta_T(\theta) = P_\theta\left(T(\mathbf{X}) < c\right) \leq P_\theta\left(\lambda(\mathbf{X}) < c\right) = \beta_\lambda(\theta),$$

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

Example 8.3.22 (An equivalence) In some situations, $T(\mathbf{x}) = \lambda(\mathbf{x})$ in Theorem 8.3.21. The UIT built up from individual LRTs is the same as the overall LRT. This was the case in Example 8.2.8. There the UIT formed from two one-sided t tests was equivalent to the two-sided LRT. ||

Since the LRT is uniformly more powerful than the UIT in Theorem 8.3.21, 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. As yet, we have not discussed inferences for the individual $H_{0\gamma}$. The error probabilities for such inferences would have to be examined before such an inference procedure were adopted. But 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. A simple bound for the size of an IUT is related to the sizes of the individual tests that are used to define the IUT. Recall that in this situation the null hypothesis is expressible as a *union*; that is, we are testing

$$H_0\colon \theta \in \Theta_0 \qquad \text{versus} \qquad H_1\colon \theta \in \Theta_0^c, \quad \text{where } \Theta_0 = \bigcup_{\gamma\in\Gamma}\Theta_\gamma.$$

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 $H_{0\gamma}\colon \theta \in \Theta_\gamma$.

Theorem 8.3.23 *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

$$P_\theta(\mathbf{X} \in R) \leq P_\theta(\mathbf{X} \in R_\gamma) \leq \alpha_\gamma \leq \alpha.$$

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

Typically, the individual rejection regions R_γ are chosen so that $\alpha_\gamma = \alpha$ for all γ. In such a case, Theorem 8.3.23 states that the resulting IUT is a level α test.

Theorem 8.3.23, which provides an upper bound for the size of an IUT, is somewhat more useful than Theorem 8.3.21, which provides an upper bound for the size of a UIT. Theorem 8.3.21 applies only to UITs constructed from likelihood ratio tests. In contrast, Theorem 8.3.23 applies to any IUT.

The bound in Theorem 8.3.21 is the size of the LRT, which, in a complicated problem, may be difficult to compute. In Theorem 8.3.23, 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 $\alpha_\gamma, \gamma \in \Gamma$.

The IUT in Theorem 8.3.23 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 size of the IUT is exactly α and the IUT is not too conservative.

Theorem 8.3.24 *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. $\lim_{l\to\infty} P_{\theta_l}(\mathbf{X} \in R_i) = \alpha$,

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

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

Proof: By Theorem 8.3.23, R is a level α test, that is,

$$\sup_{\theta\in\Theta_0} P_\theta(\mathbf{X} \in R) \leq \alpha. \tag{8.3.7}$$

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

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

This and (8.3.7) imply the test has size exactly equal to α. □

Example 8.3.25 (Intersection–union test) In Example 8.2.9, let $n = m = 58$, $t = 1.672$, and $b = 57$. Then each of the individual tests has size $\alpha = .05$ (approximately). Therefore, by Theorem 8.3.23, the IUT is a level $\alpha = .05$ test; that is, the probability of deciding the product is good, when in fact it is not, is no more than .05. In fact, this test is a size $\alpha = .05$ test. To see this consider a sequence of parameter points $\theta_l = (\theta_{1l}, \theta_2)$, with $\theta_{1l} \to \infty$ as $l \to \infty$ and $\theta_2 = .95$. All such parameter points are in Θ_0 because $\theta_2 \leq .95$. Also, $P_{\theta_l}(\mathbf{X} \in R_1) \to 1$ as $\theta_{1l} \to \infty$, while $P_{\theta_l}(\mathbf{X} \in R_2) = .05$ for all l because $\theta_2 = .95$. Thus, by Theorem 8.3.24, 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 $Y_1, \ldots, Y_m$ 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. For example, X_i and Y_i may be related if they are measurements on the same piece of fabric, but this relationship would have to be modeled and used to calculate the exact power of the IUT at any particular parameter value.

8.3.4 p-Values

After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way. One method of reporting the results of a hypothesis test is to report the size, α, of the test used and the decision to reject H_0 or accept H_0. The size of the test carries important information. If α is small, the decision to reject H_0 is fairly convincing, but if α is large, the decision to reject H_0 is not very convincing because the test has a large probability of incorrectly making that decision. Another way of reporting the results of a hypothesis test is to report the value of a certain kind of test statistic called a *p-value.*

Definition 8.3.26 A *p-value* $p(\mathbf{X})$ is a test statistic satisfying $0 \le p(\mathbf{x}) \le 1$ for every sample point $\mathbf{x}$. Small values of $p(\mathbf{X})$ give evidence that H_1 is true. A p-value is *valid* if, for every $\theta \in \Theta_0$ and every $0 \le \alpha \le 1$,

$$P_\theta\left(p(\mathbf{X}) \le \alpha\right) \le \alpha. \tag{8.3.8}$$

If $p(\mathbf{X})$ is a valid p-value, it is easy to construct a level α test based on $p(\mathbf{X})$. The test that rejects H_0 if and only if $p(\mathbf{X}) \le \alpha$ is a level α test because of (8.3.8). 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 $p(\mathbf{x})$ 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 most common way to define a valid p-value is given in Theorem 8.3.27.

Theorem 8.3.27 *Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define*

$$p(\mathbf{x}) = \sup_{\theta\in\Theta_0} P_\theta\left(W(\mathbf{X}) \ge W(\mathbf{x})\right). \tag{8.3.9}$$

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

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

$$p_\theta(\mathbf{x}) = P_\theta\left(W(\mathbf{X}) \ge W(\mathbf{x})\right) = P_\theta\left(-W(\mathbf{X}) \le -W(\mathbf{x})\right) = F_\theta\left(-W(\mathbf{x})\right).$$

Then the random variable $p_\theta(\mathbf{X})$ is equal to $F_\theta(-W(\mathbf{X}))$. Hence, by the Probability Integral Transformation or Exercise 2.10, the distribution of $p_\theta(\mathbf{X})$ is stochastically greater than or equal to a uniform(0, 1) distribution. That is, for every $0 \le \alpha \le 1$, $P_\theta(p_\theta(\mathbf{X}) \le \alpha) \le \alpha$. Because $p(\mathbf{x}) = \sup_{\theta' \in \Theta_0} p_{\theta'}(\mathbf{x}) \ge p_\theta(\mathbf{x})$ for every $\mathbf{x}$,

$$P_\theta\left(p(\mathbf{X}) \le \alpha\right) \le P_\theta\left(p_\theta(\mathbf{X}) \le \alpha\right) \le \alpha.$$

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

The calculation of the supremum in (8.3.9) might be difficult. The next two examples illustrate common situations in which it is not too difficult. In the first, no supremum is necessary; in the second, it is easy to determine the θ value at which the supremum occurs.

Example 8.3.28 (Two-sided normal p-value) Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population. Consider testing $H_0\colon \mu = \mu_0$ versus $H_1\colon \mu \ne \mu_0$. By Exercise 8.38, the LRT rejects H_0 for large values of $W(\mathbf{X}) = |\bar{X} - \mu_0|/(S/\sqrt{n})$. 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 (8.3.9), the probability is the same for all values of θ, that is, all values of σ. Thus, the p-value from (8.3.9) for this two-sided t test is $p(\mathbf{x}) = 2P(T_{n-1} \ge |\bar{x} - \mu_0|/(s/\sqrt{n}))$, where T_{n-1} has a Student's t distribution with $n-1$ degrees of freedom. ||

Example 8.3.29 (One-sided normal p-value) Again consider the normal model of Example 8.3.28, but consider testing $H_0\colon \mu \le \mu_0$ versus $H_1\colon \mu > \mu_0$. By Exercise 8.37, the LRT rejects H_0 for large values of $W(\mathbf{X}) = (\bar{X} - \mu_0)/(S/\sqrt{n})$. The following argument shows that, for this statistic, the supremum in (8.3.9) always occurs at a parameter (μ_0, σ), and the value of σ used does not matter. Consider any $\mu \le \mu_0$ and any σ:

$$\begin{aligned}
P_{\mu,\sigma}\left(W(\mathbf{X}) \ge W(\mathbf{x})\right) &= P_{\mu,\sigma}\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \ge W(\mathbf{x})\right) \\
&= P_{\mu,\sigma}\left(\frac{\bar{X} - \mu}{S/\sqrt{n}} \ge W(\mathbf{x}) + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) \\
&= P_{\mu,\sigma}\left(T_{n-1} \ge W(\mathbf{x}) + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) \\
&\le P\left(T_{n-1} \ge W(\mathbf{x})\right).
\end{aligned}$$

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 $\mu_0 \ge \mu$ and $(\mu_0 - \mu)/(S/\sqrt{n})$ is a nonnegative random variable. The subscript on P is dropped here, because this probability does not depend on (μ, σ). Furthermore,

$$P\left(T_{n-1} \ge W(\mathbf{x})\right) = P_{\mu_0,\sigma}\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \ge W(\mathbf{x})\right) = P_{\mu_0,\sigma}\left(W(\mathbf{X}) \ge W(\mathbf{x})\right),$$

and this probability is one of those considered in the calculation of the supremum in (8.3.9) because $(\mu_0, \sigma) \in \Theta_0$. Thus, the p-value from (8.3.9) for this one-sided t test is $p(\mathbf{x}) = P(T_{n-1} \ge W(\mathbf{x})) = P(T_{n-1} \ge (\bar{x} - \mu_0)/(s/\sqrt{n}))$. ||

Another method for defining a valid p-value, an alternative to using (8.3.9), involves conditioning on a sufficient statistic. Suppose $S(\mathbf{X})$ is a sufficient statistic for the model $\{f(\mathbf{x}|\theta) : \theta \in \Theta_0\}$. (To avoid tests with low power it is important that S is sufficient only for the null model, not the entire model $\{f(\mathbf{x}|\theta) : \theta \in \Theta\}$.) If the null hypothesis is true, the conditional distribution of $\mathbf{X}$ given $S = s$ does not depend on θ. Again, let $W(\mathbf{X})$ denote a test statistic for which large values give evidence that H_1 is true. Then, for each sample point $\mathbf{x}$ define

$$p(\mathbf{x}) = P\left(W(\mathbf{X}) \geq W(\mathbf{x}) | S = S(\mathbf{x})\right). \tag{8.3.10}$$

Arguing as in Theorem 8.3.27, but considering only the single distribution that is the conditional distribution of $\mathbf{X}$ given $S = s$, we see that, for any $0 \leq \alpha \leq 1$,

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

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

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

Thus, $p(\mathbf{X})$ defined by (8.3.10) is a valid p-value. Sums can be replaced by integrals for continuous S, but this method is usually used for discrete S, as in the next example.

Example 8.3.30 (Fisher's Exact Test) Let S_1 and S_2 be independent observations with $S_1 \sim \text{binomial}(n_1, p_1)$ and $S_2 \sim \text{binomial}(n_2, p_2)$. 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

$$\begin{aligned} f(s_1, s_2|p) &= \binom{n_1}{s_1} p^{s_1}(1-p)^{n_1-s_1} \binom{n_2}{s_2} p^{s_2}(1-p)^{n_2-s_2} \\ &= \binom{n_1}{s_1}\binom{n_2}{s_2} p^{s_1+s_2}(1-p)^{n_1+n_2-(s_1+s_2)}. \end{aligned}$$

Thus, $S = S_1 + S_2$ 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 S_1 given $S = s$ is hypergeometric$(n_1 + n_2, n_1, s)$ (see Exercise 8.48). Thus the conditional p-value in (8.3.10) 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. ||

8.3.5 Loss Function Optimality

A decision theoretic analysis, as in Section 7.3.4, may be used to compare hypothesis tests, rather than just comparing them via their power functions. To carry out this kind of analysis, we must specify the action space and loss function for our hypothesis testing problem.

In a hypothesis testing problem, only two actions are allowable, "accept H_0" or "reject H_0." These two actions might be denoted a_0 and a_1, respectively. The action space in hypothesis testing is the two-point set $\mathcal{A} = \{a_0, a_1\}$. A decision rule $\delta(\mathbf{x})$ (a hypothesis test) is a function on $\mathcal{X}$ that takes on only two values, a_0 and a_1. The set $\{\mathbf{x}\colon \delta(\mathbf{x}) = a_0\}$ is the acceptance region for the test, and the set $\{\mathbf{x}\colon \delta(\mathbf{x}) = a_1\}$ is the rejection region, just as in Definition 8.1.3.

The loss function in a hypothesis testing problem should reflect the fact that, if $\theta \in \Theta_0$ and decision a_1 is made, or if $\theta \in \Theta_0^c$ and decision a_0 is made, a mistake has been made. But in the other two possible cases, the correct decision has been made. Since there are only two possible actions, the loss function $L(\theta, a)$ in a hypothesis testing problem is composed of only two parts. The function $L(\theta, a_0)$ is the loss incurred for various values of θ if the decision to accept H_0 is made, and $L(\theta, a_1)$ is the loss incurred for various values of θ if the decision to reject H_0 is made.

The simplest kind of loss in a testing problem is called *0–1 loss* and is defined by

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ 1 & \theta \in \Theta_0^c \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} 1 & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c. \end{cases}$$

With 0–1 loss, the value 0 is lost if a correct decision is made and the value 1 is lost if an incorrect decision is made. This is a particularly simple situation in which both types of error have the same consequence. A slightly more realistic loss, one that gives different costs to the two types of error, is *generalized 0–1 loss*,

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ c_{\mathrm{II}} & \theta \in \Theta_0^c \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} c_{\mathrm{I}} & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c. \end{cases} \tag{8.3.11}$$

In this loss, c_{I} is the cost of a Type I Error, the error of falsely rejecting H_0, and c_{II} is the cost of a Type II Error, the error of falsely accepting H_0. (Actually, when we compare tests, all that really matters is the ratio $c_{\mathrm{II}}/c_{\mathrm{I}}$, not the two individual values. If $c_{\mathrm{I}} = c_{\mathrm{II}}$, we essentially have 0–1 loss.)

In a decision theoretic analysis, the risk function (the expected loss) is used to evaluate a hypothesis testing procedure. The risk function of a test is closely related to its power function, as the following analysis shows.

Let $\beta(\theta)$ be the power function of the test based on the decision rule δ. That is, if $R = \{\mathbf{x}\colon \delta(\mathbf{x}) = a_1\}$ denotes the rejection region of the test, then

$$\beta(\theta) = P_\theta(\mathbf{X} \in R) = P_\theta(\delta(\mathbf{X}) = a_1).$$

The risk function associated with (8.3.11) and, in particular, 0–1 loss is very simple. For any value of $\theta \in \Theta, L(\theta, a)$ takes on only two values, 0 and c_{I} if $\theta \in \Theta_0$ and 0 and c_{II} if $\theta \in \Theta_0^c$. Thus the risk is

$$
(8.3.12)\quad
\begin{aligned}
R(\theta,\delta) &= 0P_\theta(\delta(\mathbf{X}) = a_0) + c_{\mathrm{I}}P_\theta(\delta(\mathbf{X}) = a_1) = c_{\mathrm{I}}\beta(\theta) \quad \text{if } \theta \in \Theta_0,\\
R(\theta,\delta) &= c_{\mathrm{II}}P_\theta(\delta(\mathbf{X}) = a_0) + 0P_\theta(\delta(\mathbf{X}) = a_1) = c_{\mathrm{II}}(1-\beta(\theta)) \quad \text{if } \theta \in \Theta_0^c.
\end{aligned}
$$

This similarity between a decision theoretic approach and a more traditional power approach is due, in part, to the form of the loss function. But in all hypothesis testing problems, as we shall see below, the power function plays an important role in the risk function.

Example 8.3.31 (Risk of UMP test) Let $X_1,\ldots,X_n$ be a random sample from a $\mathrm{n}(\mu,\sigma^2)$ population, σ^2 known. The UMP level α test of H_0: $\theta \geq \theta_0$ versus H_1: $\theta < \theta_0$ is the test that rejects H_0 if $(\bar{X}-\theta_0)/(\sigma/\sqrt{n}) < -z_\alpha$ (see Example 8.3.15). The power function for this test is

$$\beta(\theta) = P_\theta\left(Z < -z_\alpha - \frac{\theta-\theta_0}{\sigma/\sqrt{n}}\right),$$

where Z has a $\mathrm{n}(0,1)$ distribution. For $\alpha = .10$, the risk function (8.3.12) for $c_{\mathrm{I}} = 8$ and $c_{\mathrm{II}} = 3$ is shown in Figure 8.3.4. Notice the discontinuity in the risk function at $\theta = \theta_0$. This is due to the fact that at θ_0 the expression in the risk function changes from $\beta(\theta)$ to $1-\beta(\theta)$ as well as to the difference between c_{I} and c_{II}. ||

The 0–1 loss judges only whether a decision is right or wrong. It may be the case that some wrong decisions are more serious than others and the loss function should reflect this. When we test H_0: $\theta \geq \theta_0$ versus H_1: $\theta < \theta_0$, it is a Type I Error to reject H_0 if θ is slightly bigger than θ_0, but it may not be a very serious mistake. The adverse consequences of rejecting H_0 may be much worse if θ is much larger than θ_0. A loss function that reflects this is

$$(8.3.13)\qquad L(\theta,a_0) = \begin{cases} 0 & \theta \geq \theta_0 \\ b(\theta_0-\theta) & \theta < \theta_0 \end{cases} \quad\text{and}\quad L(\theta,a_1) = \begin{cases} c(\theta-\theta_0)^2 & \theta \geq \theta_0 \\ 0 & \theta < \theta_0, \end{cases}$$

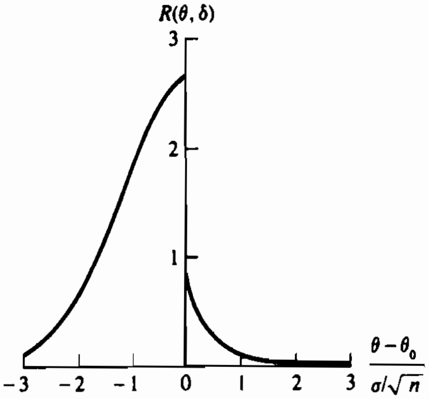

Figure 8.3.4. *Risk function for test in Example 8.3.31*

where b and c are positive constants. For example, if an experimenter is testing whether a drug lowers cholesterol level, H_0 and H_1 might be set up like this with $\theta_0 =$ standard acceptable cholesterol level. Since a high cholesterol level is associated with heart disease, the consequences of rejecting H_0 when θ is large are quite serious. A loss function like (8.3.13) reflects such a consequence. A similar type of loss function is advocated by Vardeman (1987).

Even for a general loss function like (8.3.13), the risk function and the power function are closely related. For any fixed value of θ, the loss is either $L(\theta, a_0)$ or $L(\theta, a_1)$. Thus the expected loss is

$$\begin{aligned} R(\theta,\delta) &= L(\theta,a_0)P_\theta(\delta(\mathbf{X})=a_0) + L(\theta,a_1)P_\theta(\delta(\mathbf{X})=a_1) \\ &= L(\theta,a_0)(1-\beta(\theta)) + L(\theta,a_1)\beta(\theta). \qquad (8.3.14) \end{aligned}$$

The power function of a test is always important when evaluating a hypothesis test. But in a decision theoretic analysis, the weights given by the loss function are also important.

8.4 Exercises

8.1 In 1,000 tosses of a coin, 560 heads and 440 tails appear. Is it reasonable to assume that the coin is fair? Justify your answer.

8.2 In a given city it is assumed that the number of automobile accidents in a given year follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has dropped?

8.3 Here, the LRT alluded to in Example 8.2.9 will be derived. Suppose that we observe m iid Bernoulli(θ) random variables, denoted by $Y_1, \ldots, Y_m$. Show that the LRT of $H_0\colon \theta \le \theta_0$ versus $H_1\colon \theta > \theta_0$ will reject H_0 if $\sum_{i=1}^m Y_i > b$.

8.4 Prove the assertion made in the text after Definition 8.2.1. If $f(x|\theta)$ is the pmf of a discrete random variable, then the numerator of $\lambda(\mathbf{x})$, the LRT statistic, is the maximum probability of the observed sample when the maximum is computed over parameters in the null hypothesis. Furthermore, the denominator of $\lambda(\mathbf{x})$ is the maximum probability of the observed sample over all possible parameters.

8.5 A random sample, $X_1, \ldots, X_n$, is drawn from a Pareto population with pdf

$$f(x|\theta,\nu) = \frac{\theta\nu^\theta}{x^{\theta+1}} I_{[\nu,\infty)}(x), \quad \theta > 0, \quad \nu > 0.$$

(a) Find the MLEs of θ and ν.

(b) Show that the LRT of

$$H_0\colon \theta = 1,\ \nu \text{ unknown}, \qquad \text{versus} \qquad H_1\colon \theta \ne 1,\ \nu \text{ unknown},$$

has critical region of the form $\{\mathbf{x}\colon T(\mathbf{x}) \le c_1 \text{ or } T(\mathbf{x}) \ge c_2\}$, where $0 < c_1 < c_2$ and

$$T = \log\left[\frac{\prod_{i=1}^n X_i}{(\min_i X_i)^n}\right].$$

(c) Show that, under H_0, $2T$ has a chi squared distribution, and find the number of degrees of freedom. (*Hint:* Obtain the joint distribution of the $n-1$ nontrivial terms $X_i/(\min_i X_i)$ conditional on $\min_i X_i$. Put these $n-1$ terms together, and notice that the distribution of T given $\min_i X_i$ does not depend on $\min_i X_i$, so it is the unconditional distribution of T.)

8.6 Suppose that we have two independent random samples: $X_1, \ldots, X_n$ are exponential(θ), and $Y_1, \ldots, Y_m$ are exponential(μ).

(a) Find the LRT of $H_0\colon \theta = \mu$ versus $H_1 : \theta \neq \mu$.

(b) Show that the test in part (a) can be based on the statistic

$$T = \frac{\Sigma X_i}{\Sigma X_i + \Sigma Y_i}.$$

(c) Find the distribution of T when H_0 is true.

8.7 We have already seen the usefulness of the LRT in dealing with problems with nuisance parameters. We now look at some other nuisance parameter problems.

(a) Find the LRT of

$$H_0 : \theta \leq 0 \qquad \text{versus} \qquad H_1 : \theta > 0$$

based on a sample $X_1, \ldots, X_n$ from a population with probability density function $f(x|\theta, \lambda) = \frac{1}{\lambda} e^{-(x-\theta)/\lambda} I_{[\theta,\infty)}(x)$, where both θ and λ are unknown.

(b) We have previously seen that the exponential pdf is a special case of a gamma pdf. Generalizing in another way, the exponential pdf can be considered as a special case of the Weibull(γ, β). The Weibull pdf, which reduces to the exponential if $\gamma = 1$, is very important in modeling reliability of systems. Suppose that $X_1, \ldots, X_n$ is a random sample from a Weibull population with both γ and β unknown. Find the LRT of $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$.

8.8 A special case of a normal family is one in which the mean and the variance are related, the n$(\theta, a\theta)$ family. If we are interested in testing this relationship, regardless of the value of θ, we are again faced with a nuisance parameter problem.

(a) Find the LRT of $H_0 : a = 1$ versus $H_1 : a \neq 1$ based on a sample $X_1, \ldots, X_n$ from a n$(\theta, a\theta)$ family, where θ is unknown.

(b) A similar question can be asked about a related family, the n$(\theta, a\theta^2)$ family. Thus, if $X_1, \ldots, X_n$ are iid n$(\theta, a\theta^2)$, where θ is unknown, find the LRT of $H_0 : a = 1$ versus $H_1 : a \neq 1$.

8.9 Stefanski (1996) establishes the arithmetic-geometric-harmonic mean inequality (see Example 4.7.8 and Miscellanea 4.9.2) using a proof based on likelihood ratio tests. Suppose that $Y_1, \ldots, Y_n$ are independent with pdfs $\lambda_i e^{-\lambda_i y_i}$, and we want to test H_0: $\lambda_1 = \cdots = \lambda_n$ vs. H_1: λ_i are not all equal.

(a) Show that the LRT statistic is given by $(\bar{Y})^{-n}/(\prod_i Y_i)^{-1}$ and hence deduce the arithmetic-geometric mean inequality.

(b) Make the transformation $X_i = 1/Y_i$, and show that the LRT statistic based on $X_1, \ldots, X_n$ is given by $[n/\sum_i (1/X_i)]^n / \prod_i X_i$ and hence deduce the geometric-harmonic mean inequality.

8.10 Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let λ have a gamma(α, β) distribution, the conjugate family for the Poisson. In Exercise 7.24 the posterior distribution of λ was found, including the posterior mean and variance. Now consider a Bayesian test of $H_0 : \lambda \leq \lambda_0$ versus $H_1 : \lambda > \lambda_0$.

(a) Calculate expressions for the posterior probabilities of H_0 and H_1.

(b) If $\alpha = \frac{5}{2}$ and $\beta = 2$, the prior distribution is a chi squared distribution with 5 degrees of freedom. Explain how a chi squared table could be used to perform a Bayesian test.

8.11 In Exercise 7.23 the posterior distribution of σ^2, the variance of a normal population, given S^2, the sample variance based on a sample of size n, was found using a conjugate prior for σ^2 (the inverted gamma pdf with parameters α and β). Based on observing S^2, a decision about the hypotheses $H_0\colon \sigma \le 1$ versus $H_1 : \sigma > 1$ is to be made.

(a) Find the region of the sample space for which $P(\sigma \le 1|s^2) > P(\sigma > 1|s^2)$, the region for which a Bayes test will decide that $\sigma \le 1$.

(b) Compare the region in part (a) with the acceptance region of an LRT. Is there any choice of prior parameters for which the regions agree?

8.12 For samples of size $n = 1, 4, 16, 64, 100$ from a normal population with mean μ and known variance σ^2, plot the power function of the following LRTs. Take $\alpha = .05$.

(a) $H_0\colon \mu \le 0$ versus $H_1 : \mu > 0$

(b) $H_0\colon \mu = 0$ versus $H_1 : \mu \ne 0$

8.13 Let X_1, X_2 be iid uniform$(\theta, \theta + 1)$. For testing $H_0\colon \theta = 0$ versus $H_1 : \theta > 0$, we have two competing tests:

$$\phi_1(X_1) \text{ :Reject } H_0 \text{ if } X_1 > .95,$$

$$\phi_2(X_1, X_2) \text{ :Reject } H_0 \text{ if } X_1 + X_2 > C.$$

(a) Find the value of C so that ϕ_2 has the same size as ϕ_1.

(b) Calculate the power function of each test. Draw a well-labeled graph of each power function.

(c) Prove or disprove: ϕ_2 is a more powerful test than ϕ_1.

(d) Show how to get a test that has the same size but is more powerful than ϕ_2.

8.14 For a random sample $X_1, \ldots, X_n$ of Bernoulli(p) variables, it is desired to test

$$H_0\colon p = .49 \qquad \text{versus} \qquad H_1 : p = .51.$$

Use the Central Limit Theorem to determine, approximately, the sample size needed so that the two probabilities of error are both about .01. Use a test function that rejects H_0 if $\sum_{i=1}^n X_i$ is large.

8.15 Show that for a random sample $X_1, \ldots, X_n$ from a n$(0, \sigma^2)$ population, the most powerful test of $H_0\colon \sigma = \sigma_0$ versus $H_1 : \sigma = \sigma_1$, where $\sigma_0 < \sigma_1$, is given by

$$\phi(\Sigma X_i^2) = \begin{cases} 1 & \text{if } \Sigma X_i^2 > c \\ 0 & \text{if } \Sigma X_i^2 \le c. \end{cases}$$

For a given value of α, the size of the Type I Error, show how the value of c is explicitly determined.

8.16 One very striking abuse of α levels is to choose them *after* seeing the data and to choose them in such a way as to force rejection (or acceptance) of a null hypothesis. To see what the *true* Type I and Type II Error probabilities of such a procedure are, calculate size and power of the following two trivial tests:

(a) Always reject H_0, no matter what data are obtained (equivalent to the practice of choosing the α level to force rejection of H_0).

(b) Always accept H_0, no matter what data are obtained (equivalent to the practice of choosing the α level to force acceptance of H_0).

8.17 Suppose that $X_1, \ldots, X_n$ are iid with a beta$(\mu, 1)$ pdf and $Y_1, \ldots, Y_m$ are iid with a beta$(\theta, 1)$ pdf. Also assume that the Xs are independent of the Ys.

(a) Find an LRT of $H_0\colon \theta = \mu$ versus $H_1\colon \theta \neq \mu$.

(b) Show that the test in part (a) can be based on the statistic

$$T = \frac{\Sigma \log X_i}{\Sigma \log X_i + \Sigma \log Y_i}.$$

(c) Find the distribution of T when H_0 is true, and then show how to get a test of size $\alpha = .10$.

8.18 Let $X_1, \ldots, X_n$ be a random sample from a n(θ, σ^2) population, σ^2 known. An LRT of $H_0\colon \theta = \theta_0$ versus $H_1\colon \theta \neq \theta_0$ is a test that rejects H_0 if $|\bar{X} - \theta_0|/(\sigma/\sqrt{n}) > c$.

(a) Find an expression, in terms of standard normal probabilities, for the power function of this test.

(b) The experimenter desires a Type I Error probability of .05 and a maximum Type II Error probability of .25 at $\theta = \theta_0 + \sigma$. Find values of n and c that will achieve this.

8.19 The random variable X has pdf $f(x) = e^{-x}, x > 0$. One observation is obtained on the random variable $Y = X^\theta$, and a test of $H_0 : \theta = 1$ versus $H_1 : \theta = 2$ needs to be constructed. Find the UMP level $\alpha = .10$ test and compute the Type II Error probability.

8.20 Let X be a random variable whose pmf under H_0 and H_1 is given by

x	1	2	3	4	5	6	7
$f(x\|H_0)$	.01	.01	.01	.01	.01	.01	.94
$f(x\|H_1)$	.06	.05	.04	.03	.02	.01	.79

Use the Neyman–Pearson Lemma to find the most powerful test for H_0 versus H_1 with size $\alpha = .04$. Compute the probability of Type II Error for this test.

8.21 In the proof of Theorem 8.3.12 (Neyman–Pearson Lemma), it was stated that the proof, which was given for continuous random variables, can easily be adapted to cover discrete random variables. Provide the details; that is, prove the Neyman–Pearson Lemma for discrete random variables. Assume that the α level is attainable.

8.22 Let $X_1, \ldots, X_{10}$ be iid Bernoulli(p).

(a) Find the most powerful test of size $\alpha = .0547$ of the hypotheses $H_0\colon p = \frac{1}{2}$ versus $H_1\colon p = \frac{1}{4}$. Find the power of this test.

(b) For testing $H_0 : p \leq \frac{1}{2}$ versus $H_1 : p > \frac{1}{2}$, find the size and sketch the power function of the test that rejects H_0 if $\sum_{i=1}^{10} X_i \geq 6$.

(c) For what α levels does there exist a UMP test of the hypotheses in part (a)?

8.23 Suppose X is one observation from a population with beta$(\theta, 1)$ pdf.

(a) For testing $H_0\colon \theta \leq 1$ versus $H_1\colon \theta > 1$, find the size and sketch the power function of the test that rejects H_0 if $X > \frac{1}{2}$.

(b) Find the most powerful level α test of $H_0\colon \theta = 1$ versus $H_1\colon \theta = 2$.

(c) Is there a UMP test of $H_0\colon \theta \leq 1$ versus $H_1\colon \theta > 1$? If so, find it. If not, prove so.

8.24 Find the LRT of a *simple* H_0 versus a *simple* H_1. Is this test equivalent to the one obtained from the Neyman–Pearson Lemma? (This relationship is treated in some detail by Solomon 1975.)

8.25 Show that each of the following families has an MLR.

(a) $\mathrm{n}(\theta, \sigma^2)$ family with σ^2 known

(b) Poisson(θ) family

(c) binomial(n, θ) family with n known

8.26 (a) Show that if a family of pdfs $\{f(x|\theta)\colon \theta \in \Theta\}$ has an MLR, then the corresponding family of cdfs is *stochastically increasing* in θ. (See the Miscellanea section.)

(b) Show that the converse of part (a) is false; that is, give an example of a family of cdfs that is stochastically increasing in θ for which the corresponding family of pdfs does not have an MLR.

8.27 Suppose $g(t|\theta) = h(t)c(\theta)e^{w(\theta)t}$ is a one-parameter exponential family for the random variable T. Show that this family has an MLR if $w(\theta)$ is an increasing function of θ. Give three examples of such a family.

8.28 Let $f(x|\theta)$ be the logistic location pdf

$$f(x|\theta) = \frac{e^{(x-\theta)}}{\left(1 + e^{(x-\theta)}\right)^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

(a) Show that this family has an MLR.

(b) Based on one observation, X, find the most powerful size α test of $H_0 : \theta = 0$ versus $H_1 : \theta = 1$. For $\alpha = .2$, find the size of the Type II Error.

(c) Show that the test in part (b) is UMP size α for testing $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. What can be said about UMP tests in general for the logistic location family?

8.29 Let X be one observation from a Cauchy(θ) distribution.

(a) Show that this family does not have an MLR.

(b) Show that the test

$$\phi(x) = \begin{cases} 1 & \text{if } 1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

is most powerful of its size for testing $H_0 : \theta = 0$ versus $H_1 : \theta = 1$. Calculate the Type I and Type II Error probabilities.

(c) Prove or disprove: The test in part (b) is UMP for testing $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. What can be said about UMP tests in general for the Cauchy location family?

8.30 Let $f(x|\theta)$ be the Cauchy *scale* pdf

$$f(x|\theta) = \frac{\theta}{\pi} \frac{1}{\theta^2 + x^2}, \quad -\infty < x < \infty, \quad \theta > 0.$$

(a) Show that this family does not have an MLR.

(b) If X is one observation from $f(x|\theta)$, show that $|X|$ is sufficient for θ and that the distribution of $|X|$ *does* have an MLR.

8.31 Let $X_1, \ldots, X_n$ be iid Poisson(λ).

(a) Find a UMP test of $H_0 : \lambda \leq \lambda_0$ versus $H_1 : \lambda > \lambda_0$.

(b) Consider the specific case $H_0 : \lambda \leq 1$ versus $H_1 : \lambda > 1$. Use the Central Limit Theorem to determine the sample size n so a UMP test satisfies $P(\text{reject } H_0 | \lambda = 1) = .05$ and $P(\text{reject } H_0 | \lambda = 2) = .9$.

8.32 Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, 1)$, and let θ_0 be a specified value of θ.

(a) Find the UMP, size α, test of $H_0\colon \theta \geq \theta_0$ versus $H_1\colon \theta < \theta_0$.

(b) Show that there does not exist a UMP, size α, test of $H_0\colon \theta = \theta_0$ versus $H_1\colon \theta \neq \theta_0$.

8.33 Let $X_1, \ldots, X_n$ be a random sample from the uniform$(\theta, \theta+1)$ distribution. To test $H_0\colon \theta = 0$ versus $H_1\colon \theta > 0$, use the test

$$\text{reject } H_0 \text{ if } Y_n \geq 1 \text{ or } Y_1 \geq k,$$

where k is a constant, $Y_1 = \min\{X_1, \ldots, X_n\}, Y_n = \max\{X_1, \ldots, X_n\}$.

(a) Determine k so that the test will have size α.

(b) Find an expression for the power function of the test in part (a).

(c) Prove that the test is UMP size α.

(d) Find values of n and k so that the UMP .10 level test will have power at least .8 if $\theta > 1$.

8.34 In each of the following two situations, show that for any number c, if $\theta_1 \leq \theta_2$, then

$$P_{\theta_1}(T > c) \leq P_{\theta_2}(T > c).$$

(a) θ is a location parameter in the distribution of the random variable T.

(b) The family of pdfs of $T, \{g(t|\theta)\colon \theta \in \Theta\}$, has an MLR.

8.35 The usual t distribution, as derived in Section 5.3.2, is also known as a *central t distribution.* It can be thought of as the pdf of a random variable of the form $T = \mathrm{n}(0,1)/\sqrt{\chi^2_\nu/\nu}$, where the normal and the chi squared random variables are independent. A generalization of the t distribution, the *noncentral t*, is of the form $T' = \mathrm{n}(\mu, 1)/\sqrt{\chi^2_\nu/\nu}$, where the normal and the chi squared random variables are independent and we can have $\mu \neq 0$. (We have already seen a noncentral pdf, the noncentral chi squared, in (4.4.3).) Formally, if $X \sim \mathrm{n}(\mu, 1)$ and $Y \sim \chi^2_\nu$, independent of X, then $T' = X/\sqrt{Y/\nu}$ has a noncentral t distribution with ν degrees of freedom and noncentrality parameter $\delta = \sqrt{\mu^2}$.

(a) Calculate the mean and variance of T'.

(b) The pdf of T' is given by

$$f_{T'}(t|\delta) = \frac{e^{-\delta^2/2}}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})\sqrt{\nu}} \sum_{k=0}^{\infty} \frac{(2/\nu)^{k/2}(\delta t)^k}{k!} \frac{\Gamma([\nu+k+1]/2)}{(1+(t^2/\nu))^{(\nu+k+1)/2}}.$$

Show that this pdf reduces to that of a central t if $\delta = 0$.

(c) Show that the pdf of T' has an MLR in its noncentrality parameter.

8.36 We have one observation from a beta$(1, \theta)$ population.

(a) To test $H_0\colon \theta_1 \leq \theta \leq \theta_2$ versus $H_1\colon \theta < \theta_1$ or $\theta > \theta_2$, where $\theta_1 = 1$ and $\theta_2 = 2$, a test satisfies $\mathrm{E}_{\theta_1}\phi = .5$ and $\mathrm{E}_{\theta_2}\phi = .3$. Find a test that is as good, and explain why it is as good.

(b) For testing $H_0\colon \theta = \theta_1$ versus $H_1\colon \theta \neq \theta_1$, with $\theta_1 = 1$, find a two-sided test (other than $\phi \equiv .1$) that satisfies $\mathrm{E}_{\theta_1}\phi = .1$ and $\frac{d}{d\theta}\mathrm{E}_\theta(\phi)\big|_{\theta=\theta_1} = 0$.

8.37 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population. Consider testing

$$H_0\colon \theta \leq \theta_0 \qquad \text{versus} \qquad H_1\colon \theta > \theta_0.$$

(a) If σ^2 is known, show that the test that rejects H_0 when

$$\bar{X} > \theta_0 + z_\alpha \sqrt{\sigma^2/n}$$

is a test of size α. Show that the test can be derived as an LRT.

(b) Show that the test in part (a) is a UMP test.

(c) If σ^2 is unknown, show that the test that rejects H_0 when

$$\bar{X} > \theta_0 + t_{n-1,\alpha} \sqrt{S^2/n}$$

is a test of size α. Show that the test can be derived as an LRT.

8.38 Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$, where θ_0 is a specified value of θ and σ^2 is unknown. We are interested in testing

$$H_0\colon \theta = \theta_0 \qquad \text{versus} \qquad H_1\colon \theta \neq \theta_0.$$

(a) Show that the test that rejects H_0 when

$$|\bar{X} - \theta_0| > t_{n-1,\alpha/2} \sqrt{S^2/n}$$

is a test of size α.

(b) Show that the test in part (a) can be derived as an LRT.

8.39 Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$. We are interested in testing

$$H_0\colon \mu_X = \mu_Y \qquad \text{versus} \qquad H_1\colon \mu_X \neq \mu_Y.$$

(a) Show that the random variables $W_i = X_i - Y_i$ are iid $\mathrm{n}(\mu_W, \sigma_W^2)$.

(b) Show that the above hypothesis can be tested with the statistic

$$T_W = \frac{\bar{W}}{\sqrt{\frac{1}{n} S_W^2}},$$

where $\bar{W} = \frac{1}{n}\sum_{i=1}^n W_i$ and $S_W^2 = \frac{1}{(n-1)}\sum_{i=1}^n (W_i - \bar{W})^2$. Furthermore, show that, under H_0, $T_W \sim$ Student's t with $n-$ 1 degrees of freedom. (This test is known as the *paired-sample t test.*)

8.40 Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$.

(a) Derive the LRT of

$$H_0\colon \mu_X = \mu_Y \qquad \text{versus} \qquad H_1\colon \mu_X \neq \mu_Y,$$

where σ_X^2, σ_Y^2, and ρ are unspecified and unknown.

(b) Show that the test derived in part (a) is equivalent to the paired t test of Exercise 8.39.

(*Hint:* Straightforward maximization of the bivariate likelihood is possible but somewhat nasty. Filling in the gaps of the following argument gives a more elegant proof.)

Make the transformation $u = x - y, v = x + y$. Let $f(x, y)$ denote the bivariate normal pdf, and write

$$f(x, y) = g(v|u)h(u),$$

where $g(v|u)$ is the conditional pdf of V given U, and $h(u)$ is the marginal pdf of U. Argue that (1) the likelihood can be equivalently factored and (2) the piece involving $g(v|u)$ has the same maximum whether or not the means are restricted. Thus, it can be ignored (since it will cancel) and the LRT is based only on $h(u)$. However, $h(u)$ is a normal pdf with mean $\mu_X - \mu_Y$, and the LRT is the usual one-sample t test, as derived in Exercise 8.38.

8.41 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu_X, \sigma_X^2)$, and let $Y_1, \ldots, Y_m$ be an independent random sample from a $\mathrm{n}(\mu_Y, \sigma_Y^2)$. We are interested in testing

$$H_0\colon \mu_X = \mu_Y \qquad \text{versus} \qquad H_1\colon \mu_X \neq \mu_Y$$

with the assumption that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

(a) Derive the LRT for these hypotheses. Show that the LRT can be based on the statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_\mathrm{p}^2\left(\frac{1}{n} + \frac{1}{m}\right)}},$$

where

$$S_\mathrm{p}^2 = \frac{1}{(n+m-2)}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2\right).$$

(The quantity S_p^2 is sometimes referred to as a *pooled variance estimate.* This type of estimate will be used extensively in Section 11.2.)

(b) Show that, under H_0, $T \sim t_{n+m-2}$. (This test is known as the *two-sample t test.*)

(c) Samples of wood were obtained from the core and periphery of a certain Byzantine church. The date of the wood was determined, giving the following data.

Core		Periphery	
1294	1251	1284	1274
1279	1248	1272	1264
1274	1240	1256	1256
1264	1232	1254	1250
1263	1220	1242	
1254	1218		
1251	1210		

Use the two-sample t test to determine if the mean age of the core is the same as the mean age of the periphery.

8.42 The assumption of equal variances, which was made in Exercise 8.41, is not always tenable. In such a case, the distribution of the statistic is no longer a t. Indeed, there is doubt as to the wisdom of calculating a pooled variance estimate. (This problem, of making inference on means when variances are unequal, is, in general, quite a difficult one. It is known as the *Behrens–Fisher Problem.*) A natural test to try is the following modification of the two-sample t test: Test

$$H_0\colon \mu_X = \mu_Y \qquad \text{versus} \qquad H_1\colon \mu_X \neq \mu_Y,$$

where we do not assume that $\sigma_X^2 = \sigma_Y^2$, using the statistic

$$T' = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)}},$$

where

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{and} \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

The exact distribution of T' is not pleasant, but we can approximate the distribution using Satterthwaite's approximation (Example 7.2.3).

(a) Show that

$$\frac{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \sim \frac{\chi_\nu^2}{\nu} \qquad \text{(approximately)},$$

where ν can be estimated with

$$\hat{\nu} = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}.$$

(b) Argue that the distribution of T' can be approximated by a t distribution with $\hat{\nu}$ degrees of freedom.

(c) Re-examine the data from Exercise 8.41 using the approximate t test of this exercise; that is, test if the mean age of the core is the same as the mean age of the periphery using the T' statistic.

(d) Is there any statistical evidence that the variance of the data from the core may be different from the variance of the data from the periphery? (Recall Example 5.4.1.)

8.43 Sprott and Farewell (1993) note that in the two-sample t test, a valid t statistic can be derived as long as the ratio of variances is known. Let $X_1, \ldots, X_{n_1}$ be a sample from a $\mathrm{n}(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{n_2}$ a sample from a $\mathrm{n}(\mu_2, \rho^2\sigma^2)$, where ρ^2 is known. Show that

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n_1} + \frac{\rho^2}{n_2}}\sqrt{\frac{n_1(n_1-1)s_X^2 + n_2(n_2-1)s_Y^2/\rho^2}{n_1+n_2-2}}}$$

has Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom and $\frac{n_2 s_Y^2}{\rho^2 n_1 s_X^2}$ has an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

Sprott and Farewell also note that the t statistic is maximized at $\rho^2 = \dfrac{n_1\sqrt{n_1 - 1}s_X^2}{n_2\sqrt{n_2 - 1}s_Y^2}$, and they suggest plotting the statistic for plausible values of ρ^2, possibly those in a confidence interval.

8.44 Verify that Test 3 in Example 8.3.20 is an unbiased level α test.

8.45 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population. Consider testing

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

Let $\bar{X}_m$ denote the sample mean of the first m observations, $X_1, \ldots, X_m$, for $m = 1, \ldots, n$. If σ^2 is known, show that for each $m = 1, \ldots, n$, the test that rejects H_0 when

$$\bar{X}_m > \theta_0 + z_\alpha \sqrt{\sigma^2/m}$$

is an unbiased size α test. Graph the power function for each of these tests if $n = 4$.

8.46 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population. Consider testing

$$H_0 : \theta_1 \leq \theta \leq \theta_2 \qquad \text{versus} \qquad H_1 : \theta < \theta_1 \text{ or } \theta > \theta_2.$$

(a) Show that the test

$$\text{reject } H_0 \text{ if } \bar{X} > \theta_2 + t_{n-1,\alpha/2}\sqrt{S^2/n} \quad \text{or} \quad \bar{X} < \theta_1 - t_{n-1,\alpha/2}\sqrt{S^2/n}$$

is not a size α test.

(b) Show that, for an appropriately chosen constant k, a size α test is given by

$$\text{reject } H_0 \text{ if } |\bar{X} - \bar{\theta}| > k\sqrt{S^2/n},$$

where $\bar{\theta} = (\theta_1 + \theta_2)/2$.

(c) Show that the tests in parts (a) and (b) are unbiased of their size. (Assume that the noncentral t distribution has an MLR.)

8.47 Consider two independent normal samples with equal variances, as in Exercise 8.41. Consider testing $H_0 : \mu_X - \mu_Y \leq -\delta$ or $\mu_X - \mu_Y \geq \delta$ versus $H_1 : -\delta < \mu_X - \mu_Y < \delta$, where δ is a specified positive constant. (This is called an *equivalence testing* problem.)

(a) Show that the size α LRT of $H_0^- : \mu_X - \mu_Y \leq -\delta$ versus $H_1^- : \mu_X - \mu_Y > -\delta$ rejects H_0^- if

$$T^- = \frac{\bar{X} - \bar{Y} - (-\delta)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \geq t_{n+m-2,\alpha}.$$

(b) Find the size α LRT of $H_0^+ : \mu_X - \mu_Y \geq \delta$ versus $H_1^+ : \mu_X - \mu_Y < \delta$.

(c) Explain how the tests in (a) and (b) can be combined into a level α test of H_0 versus H_1.

(d) Show that the test in (c) is a size α test. (*Hint:* Consider $\sigma \to 0$.)

This procedure is sometimes known as the *two one-sided tests procedure* and was derived by Schuirmann (1987) (see also Westlake 1981) for the problem of testing *bioequivalence*. See also the review article by Berger and Hsu (1996) and Exercise 9.33 for a confidence interval counterpart.

8.48 Prove the assertion in Example 8.3.30 that the conditional distribution of S_1 given S is hypergeometric.

8.49 In each of the following situations, calculate the p-value of the observed data.

(a) For testing $H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$, 7 successes are observed out of 10 Bernoulli trials.

(b) For testing $H_0 : \lambda \leq 1$ versus $H_1 : \lambda > 1, X = 3$ are observed, where $X \sim$ Poisson(λ).

(c) For testing $H_0 : \lambda \leq 1$ versus $H_1 : \lambda > 1, X_1 = 3, X_2 = 5$, and $X_3 = 1$ are observed, where $X_i \sim$ Poisson(λ), independent.

8.50 Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$, σ^2 known, and let θ have a double exponential distribution, that is, $\pi(\theta) = e^{-|\theta|/a}/(2a)$, a known. A Bayesian test of the hypotheses $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$ will decide in favor of H_1 if its posterior probability is large.

(a) For a given constant K, calculate the posterior probability that $\theta > K$, that is, $P(\theta > K | x_1, \ldots, x_n, a)$.

(b) Find an expression for $\lim_{a \to \infty} P(\theta > K | x_1, \ldots, x_n, a)$.

(c) Compare your answer in part (b) to the p-value associated with the classical hypothesis test.

8.51 Here is another common interpretation of p-values. Consider a problem of testing H_0 versus H_1. Let $W(\mathbf{X})$ be a test statistic. Suppose that for each α, $0 \leq \alpha \leq 1$, a critical value c_α can be chosen so that $\{\mathbf{x} : W(\mathbf{x}) \geq c_\alpha\}$ is the rejection region of a size α test of H_0. Using this family of tests, show that the usual p-value $p(\mathbf{x})$, defined by (8.3.9), is the smallest α level at which we could reject H_0, having observed the data $\mathbf{x}$.

8.52 Consider testing $H_0 : \theta \in \bigcup_{j=1}^{k} \Theta_j$. For each $j = 1, \ldots, k$, let $p_j(\mathbf{x})$ denote a valid p-value for testing $H_{0j} : \theta \in \Theta_j$. Let $p(\mathbf{x}) = \max_{1 \leq j \leq k} p_j(\mathbf{x})$.

(a) Show that $p(\mathbf{X})$ is a valid p-value for testing H_0.

(b) Show that the α level test defined by $p(\mathbf{X})$ is the same as an α level IUT defined in terms of individual tests based on the $p_j(\mathbf{x})$s.

8.53 In Example 8.2.7 we saw an example of a one-sided Bayesian hypothesis test. Now we will consider a similar situation, but with a two-sided test. We want to test

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta \neq 0,$$

and we observe $X_1, \ldots, X_n$, a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, σ^2 known. A type of prior distribution that is often used in this situation is a mixture of a point mass on $\theta = 0$ and a pdf spread out over H_1. A typical choice is to take $P(\theta = 0) = \frac{1}{2}$, and if $\theta \neq 0$, take the prior distribution to be $\frac{1}{2}\mathrm{n}(0, \tau^2)$, where τ^2 is known.

(a) Show that the prior defined above is proper, that is, $P(-\infty < \theta < \infty) = 1$.

(b) Calculate the posterior probability that H_0 is true, $P(\theta = 0 | x_1, \ldots, x_n)$.

(c) Find an expression for the p-value corresponding to a value of $\bar{x}$.

(d) For the special case $\sigma^2 = \tau^2 = 1$, compare $P(\theta = 0 | x_1, \ldots, x_n)$ and the p-value for a range of values of $\bar{x}$. In particular,

(i) For $n = 9$, plot the p-value and posterior probability as a function of $\bar{x}$, and show that the Bayes probability is greater than the p-value for moderately large values of $\bar{x}$.

(ii) Now, for $\alpha = .05$, set $\bar{x} = Z_{\alpha/2}/\sqrt{n}$, fixing the p-value at α for all n. Show that the posterior probability at $\bar{x} = Z_{\alpha/2}/\sqrt{n}$ goes to 1 as $n \to \infty$. This is *Lindley's Paradox.*

Note that small values of $P(\theta = 0 | x_1, \ldots, x_n)$ are evidence *against* H_0, and thus this quantity is similar in spirit to a p-value. The fact that these two quantities can have very different values was noted by Lindley (1957) and is also examined by Berger and Sellke (1987). (See the Miscellanea section.)

8.54 The discrepancies between p-values and Bayes posterior probabilities are not as dramatic in the one-sided problem, as is discussed by Casella and Berger (1987) and also mentioned in the Miscellanea section. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, and suppose that the hypotheses to be tested are

$$H_0: \theta \leq 0 \qquad \text{versus} \qquad H_1: \theta > 0.$$

The prior distribution on θ is $\mathrm{n}(0, \tau^2), \tau^2$ known, which is symmetric about the hypotheses in the sense that $P(\theta \leq 0) = P(\theta > 0) = \frac{1}{2}$.

(a) Calculate the posterior probability that H_0 is true, $P(\theta \leq 0|x_1, \ldots, x_n)$.

(b) Find an expression for the p-value corresponding to a value of $\bar{x}$, using tests that reject for large values of $\bar{X}$.

(c) For the special case $\sigma^2 = \tau^2 = 1$, compare $P(\theta \leq 0|x_1, \ldots, x_n)$ and the p-value for values of $\bar{x} > 0$. Show that the Bayes probability is always greater than the p-value.

(d) Using the expression derived in parts (a) and (b), show that

$$\lim_{\tau^2 \to \infty} P(\theta \leq 0|x_1, \ldots, x_n) = \text{p-value},$$

an equality that does not occur in the two-sided problem.

8.55 Let X have a $\mathrm{n}(\theta, 1)$ distribution, and consider testing $H_0: \theta \geq \theta_0$ versus $H_1: \theta < \theta_0$. Use the loss function (8.3.13) and investigate the three tests that reject H_0 if $X < -z_\alpha + \theta_0$ for $\alpha = .1, .3$, and $.5$.

(a) For $b = c = 1$, graph and compare their risk functions.

(b) For $b = 3, c = 1$, graph and compare their risk functions.

(c) Graph and compare the power functions of the three tests to the risk functions in parts (a) and (b).

8.56 Consider testing H_0: $p \leq \frac{1}{3}$ versus H_1: $p > \frac{1}{3}$, where $X \sim \text{binomial}(5, p)$, using 0–1 loss. Graph and compare the risk functions for the following two tests. Test I rejects H_0 if $X = 0$ or 1. Test II rejects H_0 if $X = 4$ or 5.

8.57 Consider testing $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ using 0–1 loss, where $X \sim \mathrm{n}(\mu, 1)$. Let δ_c be the test that rejects H_0 if $X > c$. For every test in this problem, there is a δ_c in the class of tests $\{\delta_c, -\infty \leq c \leq \infty\}$ that has a uniformly smaller (in μ) risk function. Let δ be the test that rejects H_0 if $1 < X < 2$. Find a test δ_c that is better than δ. (Either prove that the test is better or graph the risk functions for δ and δ_c and carefully explain why the proposed test should be better.)

8.58 Consider the hypothesis testing problem and loss function given in Example 8.3.31, and let $\sigma = n = 1$. Consider tests that reject H_0 if $X < -z_\alpha + \theta_0$. Find the value of α that minimizes the maximum value of the risk function, that is, that yields a *minimax test.*

8.5 Miscellanea

8.5.1 Monotonic Power Function

In this chapter we used the property of MLR quite extensively, particularly in relation to properties of power functions of tests. The concept of *stochastic ordering* can also be used to obtain properties of power functions. (Recall that *stochastic* *ordering* has already been encountered in previous chapters, for example, in Exercises 1.49, 3.41–3.43, and 5.19. A cdf F is *stochastically greater* than a cdf G if $F(x) \leq G(x)$ for all x, with strict inequality for some x, which implies that if $X \sim F, Y \sim G$, then $P(X > x) \geq P(Y > x)$ for all x, with strict inequality for some x. In other words, F gives more probability to greater values.)

In terms of hypothesis testing, it is often the case that the distribution under the alternative is stochastically greater than under the null distribution. For example, if we have a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population and are interested in testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, it is true that all the distributions in the alternative are stochastically greater than all those in the null. Gilat (1977) uses the property of stochastic ordering, rather than MLR, to prove monotonicity of power functions under general conditions.

8.5.2 Likelihood Ratio As Evidence

The *likelihood ratio* $L(\theta_1|\mathbf{x})/L(\theta_0|\mathbf{x}) = f(\mathbf{x}|\theta_1)/f(\mathbf{x}|\theta_0)$ plays an important role in the testing of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. This ratio is equal to the LRT statistic $\lambda(\mathbf{x})$ for values of $\mathbf{x}$ that yield small values of λ. Also, the Neyman–Pearson Lemma says that the UMP level α test of H_0 versus H_1 can be defined in terms of this ratio. This likelihood ratio also has an important Bayesian interpretation. Suppose π_0 and π_1 are our prior probabilities for θ_0 and θ_1. Then, the *posterior odds in favor of* θ_1 are

$$\frac{P(\theta = \theta_1|\mathbf{x})}{P(\theta = \theta_0|\mathbf{x})} = \frac{f(\mathbf{x}|\theta_1)\pi_1/m(\mathbf{x})}{f(\mathbf{x}|\theta_0)\pi_0/m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} \cdot \frac{\pi_1}{\pi_0}.$$

π_1/π_0 are the *prior odds in favor of* θ_1. The likelihood ratio is the amount these prior odds should be adjusted, having observed the data $\mathbf{X} = \mathbf{x}$, to obtain the posterior odds. If the likelihood ratio equals 2, then the prior odds are doubled. The likelihood ratio does not depend on the prior probabilities. Thus, it is interpreted as the evidence *in the data* favoring H_1 over H_0. This kind of interpretation is discussed by Royall (1997).

8.5.3 p-Values and Posterior Probabilities

In Section 8.2.2, where Bayes tests were discussed, we saw that the posterior probability that H_0 is true is a measure of the evidence the data provide against (or for) the null hypothesis. We also saw, in Section 8.3.4, that p-values provide a measure of data-based evidence against H_0. A natural question to ask is whether these two different measures ever agree; that is, can they be *reconciled*? Berger (James, not Roger) and Sellke (1987) contended that, in the two-sided testing problem, these measures could not be reconciled, and the Bayes measure was superior. Casella and Berger (Roger 1987) argued that the two-sided Bayes problem is artificial and that in the more natural one-sided problem, the measures of evidence can be reconciled. This reconciliation makes little difference to Schervish (1996), who argues that, as measures of evidence, p-values have serious logical flaws.

8.5.4 *Confidence Set p-Values*

Berger and Boos (1994) proposed an alternative method for computing p-values. In the common definition of a p-value (Theorem 8.3.27), the "sup" is over the entire null space Θ_0. Berger and Boos proposed taking the sup over a subset of Θ_0 called C. This set $C = C(\mathbf{X})$ is determined from the data and has the property that, if $\theta \in \Theta_0$, then $P_\theta(\theta \in C(\mathbf{X})) \geq 1 - \beta$. (See Chapter 9 for a discussion of confidence sets like C.) Then the *confidence set p-value* is

$$p_C(\mathbf{x}) = \sup_{\theta \in C(\mathbf{x})} P_\theta\left(W(\mathbf{X}) \geq W(\mathbf{x})\right) + \beta.$$

Berger and Boos showed that p_C is a valid p-value.

There are two potential advantages to p_C. The computational advantage is that it may be easier to compute the sup over the smaller set C than over the larger set Θ_0. The statistical advantage is that, having observed $\mathbf{X}$, we have some idea of the value of θ; there is a good chance $\theta \in C$. It seems irrelevant to look at values of θ that do not appear to be true. The confidence set p-value looks at only those values of θ in Θ_0 that seem plausible. Berger and Boos (1994) and Silvapulle (1996) give numerous examples of confidence set p-values. Berger (1996) points out that confidence set p-values can produce tests with improved power in the problem of comparing two binomial probabilities.