Chapter 9

TESTING HYPOTHESES

9.1 Problems of Testing Hypotheses

In Example 8.3.1 on page 473, we were interested in whether or not the mean log-rainfall μ from seeded clouds was greater than some constant, specifically 4. Hypothesis testing problems are similar in nature to the decision problem of Example 8.3.1. In general, hypothesis testing concerns trying to decide whether a parameter θ lies in one subset of the parameter space or in its complement. When θ is one-dimensional, at least one of the two subsets will typically be an interval, possibly degenerate. In this section, we introduce the notation and some common methodology associated with hypothesis testing. We also demonstrate an equivalence between hypothesis tests and confidence intervals.

The Null and Alternative Hypotheses

Example 9.1.1 Rain from Seeded Clouds. In Example 8.3.1, we modeled the log-rainfalls from 26 seeded clouds as normal random variables with unknown mean μ and unknown variance σ^2. Let $\theta = (\mu, \sigma^2)$ denote the parameter vector. We are interested in whether or not $\mu > 4$. To word this in terms of the parameter vector, we are interested in whether or not θ lies in the set $\{(\mu, \sigma^2) : \mu > 4\}$. In Example 8.6.4, we calculated the probability that $\mu > 4$ as part of a Bayesian analysis. If one does not wish to do a Bayesian analysis, one must address the question of whether or not $\mu > 4$ by other means, such as those introduced in this chapter. ◄

Consider a statistical problem involving a parameter θ whose value is unknown but must lie in a certain parameter space Ω. Suppose now that Ω can be partitioned into two disjoint subsets Ω_0 and Ω_1, and the statistician is interested in whether θ lies in Ω_0 or in Ω_1.

We shall let H_0 denote the hypothesis that $\theta \in \Omega_0$ and let H_1 denote the hypothesis that $\theta \in \Omega_1$. Since the subsets Ω_0 and Ω_1 are disjoint and $\Omega_0 \cup \Omega_1 = \Omega$, exactly one of the hypotheses H_0 and H_1 must be true. The statistician must decide which of the hypotheses H_0 or H_1 appears to be true. A problem of this type, in which there are only two possible decisions, is called a problem of *testing hypotheses*. If the statistician makes the wrong decision, he might suffer a certain loss or pay a certain cost. In many problems, he will have an opportunity to observe some data before he has to make his decision, and the observed values will provide him with information about the value of θ. A procedure for deciding which hypothesis to choose is called a *test procedure* or simply a *test*.

In our discussion up to this point, we have treated the hypotheses H_0 and H_1 on an equal basis. In most problems, however, the two hypotheses are treated quite differently.

Definition 9.1.1 Null and Alternative Hypotheses/Reject. The hypothesis H_0 is called the *null hypothesis* and the hypothesis H_1 is called the *alternative hypothesis*. When performing a test, if we decide that θ lies in Ω_1, we are said to *reject* H_0. If we decide that θ lies in Ω_0, we are said not to reject H_0.

The terminology referring to the decisions in Definition 9.1.1 is asymmetric with regard to the null and alternative hypotheses. We shall return to this point later in the section.

Example 9.1.2 Egyptian Skulls. Manly (1986, p.4) reports measurements of various dimensions of human skulls found in Egypt from various time periods. These data are attributed to Thomson and Randall-Maciver (1905). One time period is approximately 4000 B.C. We might model the observed breadth measurements (in mm) of the skulls as normal random variables with unknown mean μ and variance 26. Interest might lie in how μ compares to the breadth of a modern-day skull, about 140mm. The parameter space Ω could be the positive numbers, and we could let Ω_0 be the interval $[140, \infty)$ while $\Omega_1 = (0, 140)$. In this case, we would write the null and alternative hypotheses as

$$\begin{aligned} H_0&: \quad \mu \geq 140, \\ H_1&: \quad \mu < 140. \end{aligned}$$

More realistically, we would assume that both the mean and variance of breadth measurements were unknown. That is, each measurement is a normal random variable with mean μ and variance σ^2. In this case, the parameter would be two-dimensional, for example, $\theta = (\mu, \sigma^2)$. The parameter space Ω would then be pairs of real numbers. In this case, $\Omega_0 = [140, \infty) \times (0, \infty)$ and $\Omega_1 = (0, 140) \times (0, \infty)$, since the hypotheses only concern the first coordinate μ. The hypotheses to be tested are the same as above, but now μ is only one coordinate of a two-dimensional parameter vector. We will address problems of this type in Sec. 9.5. ◀

How did we decide that the null hypothesis should be $H_0 : \mu \geq 140$ in Example 9.1.2 rather than $\mu \leq 140$? Would we be led to the same conclusion either way? We can address these issues after we introduce the possible errors that can arise in hypothesis testing (Definition 9.1.7).

Simple and Composite Hypotheses

Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where the value of the parameter θ must lie in the parameter space Ω; Ω_0 and Ω_1 are disjoint sets with $\Omega_0 \cup \Omega_1 = \Omega$; and it is desired to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \theta \in \Omega_0, \\ H_1&: \quad \theta \in \Omega_1. \end{aligned}$$

For $i = 0$ or $i = 1$, the set Ω_i may contain just a single value of θ or it might be a larger set.

Definition 9.1.2 Simple and Composite Hypotheses. If Ω_i contains just a single value of θ, then H_i is a *simple hypothesis*. If the set Ω_i contains more than one value of θ, then H_i is a *composite hypothesis*.

Under a simple hypothesis, the distribution of the observations is completely specified. Under a composite hypothesis, it is specified only that the distribution of the observations belongs to a certain class. For example, a simple null hypothesis H_0 must have the form

$$H_0: \quad \theta = \theta_0. \tag{9.1.1}$$

Definition 9.1.3 One-Sided and Two-Sided Hypotheses. Let θ be a one-dimensional parameter. *One-sided* null hypotheses are of the form $H_0 : \theta \le \theta_0$ or $H_0 : \theta \ge \theta_0$, with the corresponding one-sided alternative hypotheses being $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$. When the null hypothesis is simple, such as (9.1.1), the alternative hypothesis is usually *two-sided*, $H_1 : \theta \ne \theta_0$.

The hypotheses in Example 9.1.2 are one-sided. In Example 9.1.3 (coming up shortly), the alternative hypothesis is two-sided. One-sided and two-sided hypotheses will be discussed in more detail in Sections 9.3 and 9.4.

The Critical Region and Test Statistics

Example 9.1.3 Testing Hypotheses about the Mean of a Normal Distribution with Known Variance. Suppose that $\boldsymbol{X} = (X_1, \dots, X_n)$ is a random sample from the normal distribution with unknown mean μ and known variance σ^2. We wish to test the hypotheses

$$\begin{aligned} H_0:& \quad \mu = \mu_0, \\ H_1:& \quad \mu \ne \mu_0. \end{aligned} \tag{9.1.2}$$

It might seem reasonable to reject H_0 if $\overline{X}_n$ is far from μ_0. For example, we could choose a number c and reject H_0 if the distance from $\overline{X}_n$ to μ_0 is more than c. One way to express this is by dividing the set S of all possible data vectors $\boldsymbol{x} = (x_1, \dots, x_n)$ (the sample space) into the two sets

$$S_0 = \{\boldsymbol{x} : -c \le \overline{X}_n - \mu_0 \le c\}, \quad \text{and} \quad S_1 = S_0^C.$$

We then reject H_0 if $\boldsymbol{X} \in S_1$, and we don't reject H_0 if $\boldsymbol{X} \in S_0$. A simpler way to express the procedure is to define the statistic $T = |\overline{X}_n - \mu_0|$, and reject H_0 if $T \ge c$. ◀

In general, consider a problem in which we wish to test the following hypotheses:

$$H_0: \theta \in \Omega_0, \quad \text{and} \quad H_1: \theta \in \Omega_1. \tag{9.1.3}$$

Suppose that before the statistician has to decide which hypothesis to choose, she can observe a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ drawn from a distribution that involves the unknown parameter θ. We shall let S denote the sample space of the n-dimensional random vector $\boldsymbol{X}$. In other words, S is the set of all possible values of the random sample.

In a problem of this type, the statistician can specify a test procedure by partitioning the sample space S into two subsets. One subset S_1 contains the values of $\boldsymbol{X}$ for which she will reject H_0, and the other subset S_0 contains the values of $\boldsymbol{X}$ for which she will not reject H_0.

Definition 9.1.4 Critical Region. The set S_1 defined above is called the *critical region* of the test.

In summary, a test procedure is determined by specifying the critical region of the test. The complement of the critical region must then contain all the outcomes for which H_0 will not be rejected.

In most hypothesis-testing problems, the critical region is defined in terms of a statistic, $T = r(\boldsymbol{X})$.

Definition 9.1.5 Test Statistic/Rejection Region. Let $\boldsymbol{X}$ be a random sample from a distribution that depends on a parameter θ. Let $T = r(\boldsymbol{X})$ be a statistic, and let R be a subset of the real line. Suppose that a test procedure for the hypotheses (9.1.3) is of the form "reject H_0 if $T \in R$." Then we call T a *test statistic*, and we call R the *rejection region* of the test.

When a test is defined in terms of a test statistic T and rejection region R, as in Definition 9.1.5, the set $S_1 = \{\boldsymbol{x} : r(\boldsymbol{x}) \in R\}$ is the critical region from Definition 9.1.4.

Typically, the rejection region for a test based on a test statistic T will be some fixed interval or the outside of some fixed interval. For example, if the test rejects H_0 when $T \ge c$, the rejection region is the interval $[c, \infty)$. Once a test statistic is being used, it is simpler to express everything in terms of the test statistic rather than try to compute the critical region from Definition 9.1.4. All of the tests in the rest of this book will be based on test statistics. Indeed, most of the tests can be written in the form "reject H_0 if $T \ge c$." (Example 9.1.7 is one of the rare exceptions.)

In Example 9.1.3, the test statistic is $T = |\overline{X}_n - \mu_0|$, and the rejection region is the interval $[c, \infty)$. One can choose a test statistic using intuitive criteria, as in Example 9.1.3, or based on theoretical considerations. Some theoretical arguments are given in Sections 9.2–9.4 for choosing certain test statistics in a variety of problems involving a single parameter. Although these theoretical results provide optimal tests in the situations in which they apply, many practical problems do not satisfy the conditions required to apply these results.

Example 9.1.4 Rain from Seeded Clouds. We can formulate the problem described in Example 9.1.1 as that of testing the hypotheses $H_0 : \mu \le 4$ versus $H_1 : \mu > 4$. We could use the same test statistic as in Example 9.1.3. Alternatively, we could use the statistic $U = n^{1/2}(\overline{X}_n - 4)/\sigma'$, which looks a lot like the random variable from Eq. (8.5.1) on which confidence intervals were based. It makes sense, in this case, to reject H_0 if U is large, since that would correspond to $\overline{X}_n$ being large compared to 4. ◀

Note: Dividing Both Parameter Space and Sample Space. In the various definitions given so far, the reader needs to keep straight two different divisions. First, we divided the parameter space Ω into two disjoint subsets, Ω_0 and Ω_1. Next, we divided the sample space S into two disjoint subsets S_0 and S_1. These divisions are related to each other, but they are not the same. For one thing, the parameter space and the sample space usually are of different dimensions, so Ω_0 will necessarily be different from S_0. The relation between the two divisions is the following: If the random sample $\boldsymbol{X}$ lies in the critical region S_1, then we reject the null hypothesis Ω_0. If $\boldsymbol{X} \in S_0$, we don't reject Ω_0. We eventually learn which set S_0 or S_1 contains $\boldsymbol{X}$. We rarely learn which set Ω_0 or Ω_1 contains θ.

The Power Function and Types of Error

Let δ stand for a test procedure of the form discussed earlier in this section, either based on a critical region or based on a test statistic. The interesting probabilistic properties of δ can be summarized by computing, for each value of $\theta \in \Omega$, either the probability $\pi(\theta|\delta)$ that the test δ will reject H_0 or the probability $1 - \pi(\theta|\delta)$ that it does not reject H_0.

Definition 9.1.6 Power Function. Let δ be a test procedure. The function $\pi(\theta|\delta)$ is called the *power function* of the test δ. If S_1 denotes the critical region of δ, then the power function $\pi(\theta|\delta)$ is determined by the relation

$$\pi(\theta|\delta) = \Pr(\boldsymbol{X} \in S_1|\theta) \quad \text{for} \quad \theta \in \Omega. \tag{9.1.4}$$

If δ is described in terms of a test statistic T and rejection region R, the power function is

$$\pi(\theta|\delta) = \Pr(T \in R|\theta) \quad \text{for} \quad \theta \in \Omega. \tag{9.1.5}$$

Since the power function $\pi(\theta|\delta)$ specifies, for each possible value of the parameter θ, the probability that δ will reject H_0, it follows that the ideal power function would be one for which $\pi(\theta|\delta) = 0$ for every value of $\theta \in \Omega_0$ and $\pi(\theta|\delta) = 1$ for every value of $\theta \in \Omega_1$. If the power function of a test δ actually had these values, then regardless of the actual value of θ, δ would lead to the correct decision with probability 1. In a practical problem, however, there would seldom exist any test procedure having this ideal power function.

Example 9.1.5 Testing Hypotheses about the Mean of a Normal Distribution with Known Variance. In Example 9.1.3, the test δ is based on the test statistic $T = |\overline{X}_n - \mu_0|$ with rejection region $R = [c, \infty)$. The distribution of $\overline{X}_n$ is the normal distribution with mean μ and variance σ^2/n. The parameter is μ because we have assumed that σ^2 is known. The power function can be computed from this distribution. Let Φ denote the standard normal c.d.f. Then

$$\begin{aligned}\Pr(T \in R|\mu) &= \Pr(\overline{X}_n \geq \mu_0 + c|\mu) + \Pr(\overline{X}_n \leq \mu_0 - c|\mu)\\ &= 1 - \Phi\left(n^{1/2}\frac{\mu_0 + c - \mu}{\sigma}\right) + \Phi\left(n^{1/2}\frac{\mu_0 - c - \mu}{\sigma}\right).\end{aligned}$$

The final expression above is the power function $\pi(\mu|\delta)$. Figure 9.1 plots the power functions of three different tests with $c = 1, 2, 3$ in the specific example in which $\mu_0 = 4$, $n = 15$, and $\sigma^2 = 9$. ◀

Since the possibility of error exists in virtually every testing problem, we should consider what kinds of errors we might make. For each value of $\theta \in \Omega_0$, the decision

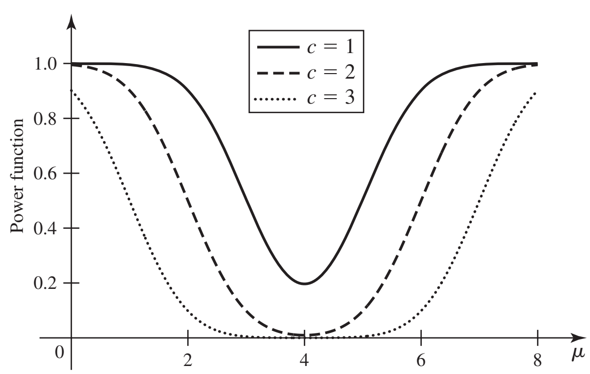


Figure 9.1 Power functions of three different tests in Example 9.1.5.

to reject H_0 is an incorrect decision. Similarly, for each value of $\theta \in \Omega_1$, the decision not to reject H_0 is an incorrect decision.

Definition 9.1.7 Type I/II Error. An erroneous decision to reject a true null hypothesis is a *type I error*, or an error of the first kind. An erroneous decision not to reject a false null hypothesis is called a *type II error*, or an error of the second kind.

In terms of the power function, if $\theta \in \Omega_0$, $\pi(\theta|\delta)$ is the probability that the statistician will make a type I error. Similarly, if $\theta \in \Omega_1$, $1 - \pi(\theta|\delta)$ is the probability of making a type II error. Of course, either $\theta \in \Omega_0$ or $\theta \in \Omega_1$, but not both. Hence, only one type of error is possible conditional on θ, but we never know which it is.

If we have our choice between several tests, we would like to choose a test δ that has small probability of error. That is, we would like the power function $\pi(\theta|\delta)$ to be low for values of $\theta \in \Omega_0$, and we would like $\pi(\theta|\delta)$ to be high for $\theta \in \Omega_1$. Generally, these two goals work against each other. That is, if we choose δ to make $\pi(\theta|\delta)$ small for $\theta \in \Omega_0$, we will usually find that $\pi(\theta|\delta)$ is small for $\theta \in \Omega_1$ as well. For example, the test procedure δ_0 that never rejects H_0, regardless of what data are observed, will have $\pi(\theta|\delta_0) = 0$ for all $\theta \in \Omega_0$. However, for this procedure $\pi(\theta|\delta_0) = 0$ for all $\theta \in \Omega_1$ as well. Similarly, the test δ_1 that always rejects H_0 will have $\pi(\theta|\delta_1) = 1$ for all $\theta \in \Omega_1$, but it will also have $\pi(\theta|\delta_1) = 1$ for all $\theta \in \Omega_0$. Hence, there is a need to strike an appropriate balance between the two goals of low power in Ω_0 and high power in Ω_1.

The most popular method for striking a balance between the two goals is to choose a number α_0 between 0 and 1 and require that

$$\pi(\theta|\delta) \le \alpha_0, \quad \text{for all } \theta \in \Omega_0. \tag{9.1.6}$$

Then, among all tests that satisfy (9.1.6), the statistician seeks a test whose power function is as high as can be obtained for $\theta \in \Omega_1$. This method is discussed in Sections 9.2 and 9.3. Another method of balancing the probabilities of type I and type II errors is to minimize a linear combination of the different probabilities of error. We shall discuss this method in Sec. 9.2 and again in Sec. 9.8.

Note: Choosing Null and Alternative Hypotheses. If one chooses to balance type I and type II error probabilities by requiring (9.1.6), then one has introduced an asymmetry in the treatment of the null and alternative hypotheses. In most testing problems, such asymmetry can be quite natural. Generally, one of the two errors (type I or type II) is more costly or less palatable in some sense. It would make sense to put tighter controls on the probability of the more serious error. For this reason, one generally arranges the null and alternative hypotheses so that type I error is the error most to be avoided. For cases in which neither hypothesis is naturally the null, switching the names of null and alternative hypotheses can have a variety of different effects on the results of testing procedures. (See Exercise 21 in this section.)

Example 9.1.6 Egyptian Skulls. In Example 9.1.2, suppose that the experimenters have a theory saying that skull breadths should increase (albeit slightly) over long periods of time. If μ is the mean breadth of skulls from 4000 B.C. and 140 is the mean breadth of modern-day skulls, the theory would say $\mu < 140$. The experimenters could mistakenly claim that the data support their theory ($\mu < 140$) when, in fact, $\mu > 140$, or they might mistakenly claim that the data fail to support their theory ($\mu > 140$) when, in fact, $\mu < 140$. In scientific studies, it is common to treat the false confirmation of one's own theory as a more serious error than falsely failing to confirm ones' own theory. This would mean type I error should be to say that $\mu < 140$ (confirm the theory, reject H_0) when, in fact, $\mu > 140$ (theory is false, H_0 is true). Traditionally, one includes the endpoints of interval hypotheses in the null, so we would formulate the hypotheses to be tested as

$$H_0:\quad \mu \ge 140,$$
$$H_1:\quad \mu < 140,$$

as we did in Example 9.1.2. ◄

The quantities in Eq. (9.1.6) play a fundamental role in hypothesis testing and have special names.

Definition 9.1.8 Level/Size. A test that satisfies (9.1.6) is called a *level α_0 test*, and we say that the test has *level of significance* α_0. In addition, the *size* $\alpha(\delta)$ of a test δ is defined as follows:

$$\alpha(\delta) = \sup_{\theta\in\Omega_0} \pi(\theta|\delta). \tag{9.1.7}$$

The following results are immediate consequences of Definition 9.1.8.

Corollary 9.1.1 A test δ is a level α_0 test if and only if its size is at most α_0 (i.e., $\alpha(\delta) \le \alpha_0$). If the null hypothesis is simple, that is, $H_0 : \theta = \theta_0$, then the size of δ will be $\alpha(\delta) = \pi(\theta_0|\delta)$. ■

Example 9.1.7 Testing Hypotheses about a Uniform Distribution. Suppose that a random sample $X_1, \ldots, X_n$ is taken from the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown ($\theta > 0$); and suppose also that it is desired to test the following hypotheses:

$$\begin{aligned} &H_0: 3 \le \theta \le 4,\\ &H_1: \theta < 3 \text{ or } \theta > 4. \end{aligned} \tag{9.1.8}$$

We know from Example 6.5.15 that the M.L.E. of θ is $Y_n = \max\{X_1, \ldots, X_n\}$. Although Y_n must be less than θ, there is a high probability that Y_n will be close to θ if the sample size n is fairly large. For illustrative purposes, suppose that the test δ does not reject H_0 if $2.9 < Y_n < 4$, and δ rejects H_0 if Y_n does not lie in this interval. Thus, the critical region of the test δ contains all the values of $X_1, \ldots, X_n$ for which either $Y_n \le 2.9$ or $Y_n \ge 4$. In terms of the test statistic Y_n, the rejection region is the union of two intervals $(-\infty, 2.9] \cup [4, \infty)$.

The power function of δ is specified by the relation

$$\pi(\theta|\delta) = \Pr(Y_n \le 2.9|\theta) + \Pr(Y_n \ge 4|\theta).$$

If $\theta \le 2.9$, then $\Pr(Y_n \le 2.9|\theta) = 1$ and $\Pr(Y_n \ge 4|\theta) = 0$. Therefore, $\pi(\theta|\delta) = 1$ if $\theta \le 2.9$. If $2.9 < \theta \le 4$, then $\Pr(Y_n \le 2.9|\theta) = (2.9/\theta)^n$ and $\Pr(Y_n \ge 4|\theta) = 0$. In this case, $\pi(\theta|\delta) = (2.9/\theta)^n$. Finally, if $\theta > 4$, then $\Pr(Y_n \le 2.9|\theta) = (2.9/\theta)^n$ and $\Pr(Y_n \ge 4|\theta) = 1 - (4/\theta)^n$. In this case, $\pi(\theta|\delta) = (2.9/\theta)^n + 1 - (4/\theta)^n$. The power function $\pi(\theta|\delta)$ is sketched in Fig. 9.2.

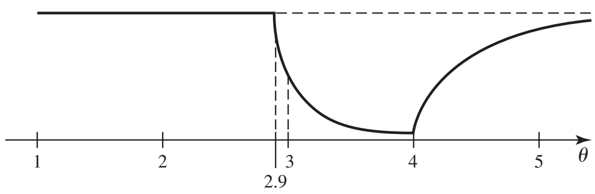
By Eq. (9.1.7), the size of δ is $\alpha(\delta) = \sup_{3\le\theta\le4} \pi(\theta|\delta)$. It can be seen from Fig. 9.2 and the calculations just given that $\alpha(\delta) = \pi(3|\delta) = (29/30)^n$. In particular, if the sample size is $n = 68$, then the size of δ is $(29/30)^{68} = 0.0997$. So δ is a level α_0 test for every level of significance $\alpha_0 \ge 0.0997$. ◄

Making a Test Have a Specific Significance Level

Suppose that we wish to test the hypotheses

$$H_0:\quad \theta \in \Omega_0,$$
$$H_1:\quad \theta \in \Omega_1.$$

Figure 9.2 The power function $\pi(\theta|\delta)$ in Example 9.1.7.



Let T be a test statistic, and suppose that our test will reject the null hypothesis if $T \geq c$, for some constant c. Suppose also that we desire our test to have the level of significance α_0. The power function of our test is $\pi(\theta|\delta) = \Pr(T \geq c|\theta)$, and we want

$$\sup_{\theta \in \Omega_0} \Pr(T \geq c|\theta) \leq \alpha_0. \tag{9.1.9}$$

It is clear that the power function, and hence the left side of (9.1.9), are nonincreasing functions of c. Hence, (9.1.9) will be satisfied for large values of c, but not for small values. If we want the power function to be as large as possible for $\theta \in \Omega_1$, we should make c as small as we can while still satisfying (9.1.9). If T has a continuous distribution, then it is usually simple to find an appropriate c.

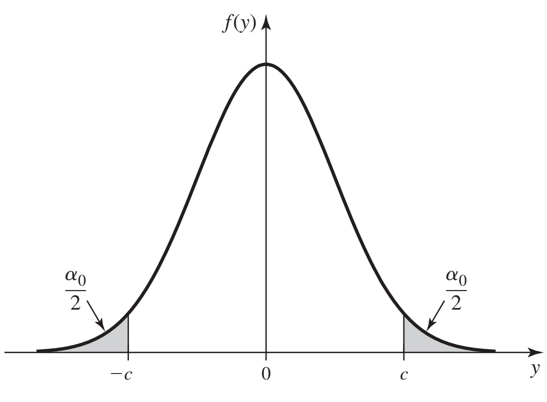
Example 9.1.8 Testing Hypotheses about the Mean of a Normal Distribution with Known Variance. In Example 9.1.5, our test is to reject $H_0: \mu = \mu_0$ if $|\overline{X}_n - \mu_0| \geq c$. Since the null hypothesis is simple, the left side of (9.1.9) reduces to the probability (assuming that $\mu = \mu_0$) that $|\overline{X}_n - \mu_0| \geq c$. Since $Y = \overline{X}_n - \mu_0$ has the normal distribution with mean 0 and variance σ^2/n when $\mu = \mu_0$, we can find a value c that makes the size exactly α_0 for each α_0. Figure 9.3 shows the p.d.f. of Y and the size of the test indicated as the shaded area under the p.d.f. Since the normal p.d.f. is symmetric around the mean (0 in this case), the two shaded areas must be the same, namely, $\alpha_0/2$. This means that c must be the $1 - \alpha_0/2$ quantile of the distribution of Y. This quantile is $c = \Phi^{-1}(1 - \alpha_0/2)\sigma n^{-1/2}$.

When testing hypotheses about the mean of a normal distribution, it is traditional to rewrite this test in terms of the statistic

$$Z = n^{1/2} \frac{\overline{X}_n - \mu_0}{\sigma}. \tag{9.1.10}$$

Then the test rejects H_0 if $|Z| \geq \Phi^{-1}(1 - \alpha_0/2)$. ◄

Figure 9.3 The p.d.f. of $Y = \overline{X}_n - \mu_0$ given $\mu = \mu_0$ for Example 9.1.8. The shaded areas represent the probability that $|Y| \geq c$.



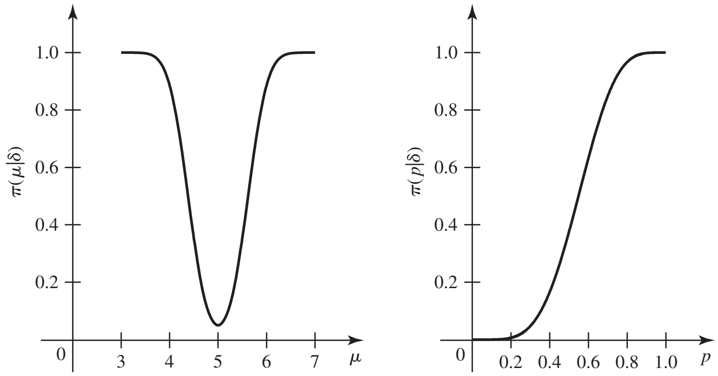


Figure 9.4 Power functions of two tests. The plot on the left is the power function of the test from Example 9.1.8 with $n = 10$, $\mu_0 = 5$, $\sigma = 1$, and $\alpha_0 = 0.05$. The plot on the right is the power function of the test from Example 9.1.9 with $n = 10$, $p_0 = 0.3$, and $\alpha_0 = 0.1$.

Example 9.1.9 Testing Hypotheses about a Bernoulli Parameter. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter p. Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad p \le p_0, \\ H_1&: \quad p > p_0. \end{aligned} \tag{9.1.11}$$

Let $Y = \sum_{i=1}^{n} X_i$, which has the binomial distribution with parameters n and p. The larger p is, the larger we expect Y to be. So, suppose that we choose to reject H_0 if $Y \ge c$, for some constant c. Suppose also that we want the size of the test to be as close to α_0 as possible without exceeding α_0. It is easy to check that $\Pr(Y \ge c|p)$ is an increasing function of p; hence, the size of the test will be $\Pr(Y \ge c|p = p_0)$. So, c should be the smallest number such that $\Pr(Y \ge c|p = p_0) \le \alpha_0$. For example, if $n = 10$, $p_0 = 0.3$, and $\alpha_0 = 0.1$, we can use the table of binomial probabilities in the back of this book to determine c. We can compute $\sum_{y=6}^{10} \Pr(Y = y|p = 0.3) = 0.0473$ and $\sum_{y=5}^{10} \Pr(Y = y|p = 0.3) = 0.1503$. In order to keep the size of the test at most 0.1, we must choose $c > 5$. Every value of c in the interval $(5, 6]$ produces the same test, since Y takes only integer values. ◀

Whenever we choose a test procedure, we should also examine the power function. If one has made a good choice, then the power function should generally be larger for $\theta \in \Omega_1$ than for $\theta \in \Omega_0$. Also, the power function should increase as θ moves away from Ω_0. For example, Fig. 9.4 shows plots of the power functions for two of the examples in this section. In both cases, the power function increases as the parameter moves away from Ω_0.

The p-value

Example 9.1.10 Testing Hypotheses about the Mean of a Normal Distribution with Known Variance. In Example 9.1.8, suppose that we choose to test the null hypothesis at level $\alpha_0 = 0.05$. We would then compute the test statistic in Eq. (9.1.10) and reject H_0 if $Z \ge \Phi^{-1}(1 - 0.05/2) = 1.96$. For example, suppose that $Z = 2.78$ is observed. Then we would reject H_0. Suppose that we were to report the result by saying that we rejected H_0 at level 0.05. What would another statistician, who felt it more appropriate to test the null hypothesis at a different level, be able to do with this report? ◀

The result of a test of hypotheses might appear to be a rather inefficient use of our data. For instance, in Example 9.1.10, we decided to reject H_0 at level $\alpha_0 = 0.05$ if the statistic Z in Eq. (9.1.10) is at least 1.96. This means that whether we observe $Z = 1.97$ or $Z = 6.97$, we shall report the same result, namely, that we rejected H_0 at level 0.05. The report of the test result does not carry any sense of how close we were to making the other decision. Furthermore, if another statistician chooses to use a size 0.01 test, then she would not reject H_0 with $Z = 1.97$, but she would reject H_0 with $Z = 6.97$. What would she do with $Z = 2.78$?

For these reasons, an experimenter does not typically choose a value of α_0 in advance of the experiment and then simply report whether or not H_0 was rejected at level α_0. In many fields of application, it has become standard practice to report, in addition to the observed value of the appropriate test statistic such as Z, *all* the values of α_0 for which the level α_0 test would lead to the rejection of H_0.

Example 9.1.11 **Testing Hypotheses about the Mean of a Normal Distribution with Known Variance.** As the observed value of Z in Example 9.1.8 is 2.78, the hypothesis H_0 would be rejected for every level of significance α_0 such that $2.78 \geq \Phi^{-1}(1 - \alpha_0/2)$. Using the table of the normal distribution given at the end of this book, this inequality translates to $\alpha_0 \geq 0.0054$. The value 0.0054 is called the *p-value* for the observed data and the tested hypotheses. Since $0.01 > 0.0054$, the statistician who wanted to test the hypotheses at level 0.01 would also reject H_0. ◀

Definition 9.1.9 ***p*-value.** In general, the *p-value* is the smallest level α_0 such that we would reject the null-hypothesis at level α_0 with the observed data.

An experimenter who rejects a null hypothesis if and only if the p-value is at most α_0 is using a test with level of significance α_0. Similarly, an experimenter who wants a level α_0 test will reject the null hypothesis if and only if the p-value is at most α_0. For this reason, the p-value is sometimes called the *observed level of significance*.

An experimenter in Example 9.1.10 would typically report that the observed value of Z was 2.78 and that the corresponding p-value was 0.0054. It is then said that the observed value of Z is *just significant* at the level of significance 0.0054. One advantage to the experimenter of reporting experimental results in this manner is that he does not need to select beforehand an arbitrary level of significance α_0 at which to carry out the test. Also, when a reader of the experimenter's report learns that the observed value of Z was just significant at the level of significance 0.0054, she immediately knows that H_0 would be rejected for every larger value of α_0 and would not be rejected for any smaller value.

Calculating p-values If all of our tests are of the form "reject the null hypothesis when $T \geq c$" for a single test statistic T, there is a straightforward way to compute p-values. For each t, let δ_t be the test that rejects H_0 if $T \geq t$. Then the p-value when $T = t$ is observed is the size of the test δ_t. (See Exercise 18.) That is, the p-value equals

$$\sup_{\theta \in \Omega_0} \pi(\theta|\delta_t) = \sup_{\theta \in \Omega_0} \Pr(T \geq t|\theta). \tag{9.1.12}$$

Typically, $\pi(\theta|\delta_t)$ is maximized at some θ_0 on the boundary between Ω_0 and Ω_1. Because the p-value is calculated as a probability in the upper tail of the distribution of T, it is sometimes called a *tail area*.

Example 9.1.12 **Testing Hypotheses about a Bernoulli Parameter.** For testing the hypotheses (9.1.11) in Example 9.1.9, we used a test that rejects H_0 if $Y \geq c$. The p-value, when $Y = y$ is observed, will be $\sup_{p \leq p_0} \Pr(Y \geq y|p)$. In this example, it is easy to see that $\Pr(Y \geq y|p)$ increases as a function of p. Hence, the p-value is $\Pr(Y \geq y|p = p_0)$. For example, let $p_0 = 0.3$ and $n = 10$. If $Y = 6$ is observed, then $\Pr(Y \geq 6|p = 0.3) = 0.0473$, as we calculated in Example 9.1.9. ◀

The calculation of the p-value is more complicated when the test cannot be put into the form "reject H_0 if $T \geq c$." In this text, we shall calculate p-values only for tests that do have this form.

Equivalence of Tests and Confidence Sets

Example 9.1.13 **Rain from Seeded Clouds.** In Examples 8.5.5 and 8.5.6, we found a coefficient γ one-sided (lower limit) confidence interval for μ, the mean log-rainfall from seeded clouds. For $\gamma = 0.9$, the observed interval is $(4.727, \infty)$. One of the controversial interpretations of this interval is that we have confidence 0.9 (whatever that means) that $\mu > 4.727$. Although this statement is deliberately ambiguous and difficult to interpret, it sounds as if it could help us address the problem of testing the hypotheses $H_0 : \mu \leq 4$ versus $H_1 : \mu > 4$. Does the fact that 4 is not in the observed coefficient 0.9 confidence interval tell us anything about whether or not we should reject H_0 at some significance level or other? ◀

We shall now illustrate how confidence intervals (see Sec. 8.5) can be used as an alternative method to report the results of a test of hypotheses. In particular, we shall show that a coefficient γ confidence set (a generalization of confidence interval to be defined shortly) can be thought of as a set of null hypotheses that would not be rejected at significance level $1 - \gamma$.

Theorem 9.1.1 **Defining Confidence Sets from Tests.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter θ. Let $g(\theta)$ be a function, and suppose that for each possible value g_0 of $g(\theta)$, there is a level α_0 test δ_{g_0} of the hypotheses

$$H_{0,g_0} : g(\theta) = g_0, \quad H_{1,g_0} : g(\theta) \neq g_0. \tag{9.1.13}$$

For each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, define

$$\omega(\boldsymbol{x}) = \{g_0 : \delta_{g_0} \text{ does not reject } H_{0,g_0} \text{ if } \boldsymbol{X} = \boldsymbol{x} \text{ is observed}\}. \tag{9.1.14}$$

Let $\gamma = 1 - \alpha_0$. Then, the random set $\omega(\boldsymbol{X})$ satisfies

$$\Pr[g(\theta_0) \in \omega(\boldsymbol{X})|\theta = \theta_0] \geq \gamma. \tag{9.1.15}$$

for all $\theta_0 \in \Omega$.

Proof Let θ_0 be an arbitrary element of Ω, and define $g_0 = g(\theta_0)$. Because δ_{g_0} is a level α_0 test, we know that

$$\Pr[\delta_{g_0} \text{ does not reject } H_{0,g_0}|\theta = \theta_0] \geq 1 - \alpha_0 = \gamma. \tag{9.1.16}$$

For each $\boldsymbol{x}$, we know that $g(\theta_0) \in \omega(\boldsymbol{x})$ if and only if the test δ_{g_0} does not reject H_{0,g_0} when $\boldsymbol{X} = \boldsymbol{x}$ is observed. It follows that the left-hand side of Eq. (9.1.15) is the same as the left-hand side of Eq. (9.1.16). ■

Definition 9.1.10 Confidence Set. If a random set $\omega(\boldsymbol{X})$ satisfies (9.1.15) for every $\theta_0 \in \Omega$, we call it a *coefficient γ confidence set for $g(\theta)$*. If the inequality in (9.1.15) is equality for all θ_0, then we call the confidence set *exact*.

A confidence set is a generalization of the concept of a confidence interval introduced in Sec. 8.5. What Theorem 9.1.1 shows is that a collection of level α_0 tests of the hypotheses (9.1.13) can be used to construct a coefficient $\gamma = 1 - \alpha_0$ confidence set for $g(\theta)$. The reverse construction is also possible.

Theorem 9.1.2 Defining Tests from Confidence Sets. Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random sample from a distribution that depends on a parameter θ. Let $g(\theta)$ be a function of θ, and let $\omega(\boldsymbol{X})$ be a coefficient γ confidence set for $g(\theta)$. For each possible value g_0 of $g(\theta)$, construct the following test δ_{g_0} of the hypotheses in Eq. (9.1.13): δ_{g_0} does not reject H_{0,g_0} if and only if $g_0 \in \omega(\boldsymbol{X})$. Then δ_{g_0} is a level $\alpha_0 = 1 - \gamma$ test of the hypotheses in Eq. (9.1.13).

Proof Because $\omega(\boldsymbol{X})$ is a coefficient γ confidence set for $g(\theta)$, it satisfies Eq. (9.1.15) for all $\theta_0 \in \Omega$. As in the proof of Theorem 9.1.1, the left-hand sides of Eqs. (9.1.15) and (9.1.16) are the same, which makes δ_{g_0} a level α_0 test. ■

Example 9.1.14 A Confidence Interval for the Mean of a Normal Distribution. Consider the test found in Example 9.1.8 for the hypotheses (9.1.2). Let $\alpha_0 = 1 - \gamma$. The size α_0 test δ_{μ_0} is to reject H_0 if $|\overline{X}_n - \mu_0| \ge \Phi^{-1}(1 - \alpha_0/2)\sigma n^{-1/2}$. If $\overline{X}_n = \overline{x}_n$ is observed, the set of μ_0 such that we would not reject H_0 is the set of μ_0 such that

$$|\overline{x}_n - \mu_0| < \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2}.$$

This inequality easily translates to

$$\overline{x}_n - \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2} < \mu_0 < \overline{x}_n + \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2}.$$

The coefficient γ confidence interval becomes

$$(A, B) = \left(\overline{X}_n - \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2}, \overline{X}_n + \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2}\right).$$

It is easy to check that $\Pr(A < \mu_0 < B | \mu = \mu_0) = \gamma$ for all μ_0. This confidence interval is exact. ◄

Example 9.1.15 Constructing a Test from a Confidence Interval. In Sec. 8.5, we learned how to construct a confidence interval for the unknown mean of a normal distribution when the variance was also unknown. Let $X_1, \dots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. In this case, the parameter is $\theta = (\mu, \sigma^2)$, and we are interested in $g(\theta) = \mu$. In Sec. 8.5, we used the statistics

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \sigma' = \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)^{1/2}. \tag{9.1.17}$$

The coefficient γ confidence interval for $g(\theta)$ is the interval

$$\left(\overline{X}_n - T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{n^{1/2}}, \overline{X}_n + T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{n^{1/2}}\right), \tag{9.1.18}$$

where $T_{n-1}^{-1}(\cdot)$ is the quantile function of the t distribution with $n-1$ degrees of freedom. For each μ_0, we can use this interval to find a level $\alpha_0 = 1-\gamma$ test of the hypotheses

$$\begin{aligned} H_0\colon \quad & \mu = \mu_0, \\ H_1\colon \quad & \mu \neq \mu_0. \end{aligned}$$

The test will reject H_0 if μ_0 is not in the interval (9.1.18). A little algebra shows that μ_0 is not in the interval (9.1.18) if and only if

$$\left| n^{1/2}\frac{\overline{X}_n - \mu_0}{\sigma'} \right| \geq T^{-1}\left(\frac{1+\gamma}{2}\right).$$

This test is identical to the t test that we shall study in more detail in Sec. 9.5. ◄

One-Sided Confidence Intervals and Tests Theorems 9.1.1 and 9.1.2 establish the equivalence between confidence sets and tests of hypotheses of the form (9.1.13). It is often necessary to test other forms of hypotheses, and it would be nice to have versions of Theorems 9.1.1 and 9.1.2 to deal with these cases. Example 9.1.13 is one such case in which the hypotheses are of the form

$$H_{0,g_0} : g(\theta) \leq g_0, \qquad H_{1,g_0} : g(\theta) > g_0. \tag{9.1.19}$$

Theorem 9.1.1 extends immediately to such cases. We leave the proof of Theorem 9.1.3 to the reader.

Theorem 9.1.3 One-Sided Confidence Intervals from One-Sided Tests. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter θ. Let $g(\theta)$ be a real-valued function, and suppose that for each possible value g_0 of $g(\theta)$, there is a level α_0 test δ_{g_0} of the hypotheses (9.1.19). For each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, define $\omega(\boldsymbol{x})$ by Eq. (9.1.14). Let $\gamma = 1-\alpha_0$. Then the random set $\omega(\boldsymbol{X})$ satisfies Eq. (9.1.15) for all $\theta_0 \in \Omega$. ■

Example 9.1.16 One-Sided Confidence Interval for a Bernoulli Parameter. In Example 9.1.9, we showed how to construct a level α_0 test of the one-sided hypotheses (9.1.11). Let $Y = \sum_{i=1}^{n} X_i$. The test rejects H_0 if $Y \geq c(p_0)$ where $c(p_0)$ is the smallest number c such that $\Pr(Y \geq c|p = p_0) \leq \alpha_0$. After observing the data $\boldsymbol{X}$, we can check, for each p_0, whether or not we reject H_0. That is, for each p_0 we check whether or not $Y \geq c(p_0)$. All those p_0 for which $Y < c(p_0)$ (i.e., we don't reject H_0) will form an interval $\omega(\boldsymbol{X})$. This interval will satisfy $\Pr(p_0 \in \omega(\boldsymbol{X})|p = p_0) \geq 1-\alpha_0$ for all p_0. For example, suppose that $n = 10$, $\alpha_0 = 0.1$, and $Y = 6$ is observed. In order not to reject $H_0 : p \leq p_0$ at level 0.1, we must have a rejection region that does not contain 6. This will happen if and only if $\Pr(Y \geq 6|p = p_0) > 0.1$. By trying various values of p_0, we find that this inequality holds for all $p_0 > 0.3542$. So, if $Y = 6$ is observed, our coefficient 0.9 confidence interval is (0.3542, 1). Notice that 0.3 is not in the interval, so we would reject $H_0 : p \leq 0.3$ with a level 0.1 test as we did in Example 9.1.9. For other observed values $Y = y$, the confidence intervals will all be of the form $(q(y), 1)$ where $q(y)$ can be computed as outlined in Exercise 17. For $n = 10$ and $\alpha_0 = 0.1$, the values of $q(y)$ are

y	0	1	2	3	4	5	6	7	8	9	10
$q(y)$	0	0.0104	0.0545	0.1158	0.1875	0.2673	0.3542	0.4482	0.5503	0.6631	0.7943

This confidence interval is not exact. ◄

Unfortunately, Theorem 9.1.2 does not immediately extend to one-sided hypotheses for the following reason. The size of a one-sided test for hypotheses of the form (9.1.19) depends on *all* of the values of θ such that $g(\theta) \le g_0$, not just on those for which $g(\theta) = g_0$. In particular, the size of the test δ_{g_0} defined in Theorem 9.1.2 is

$$\sup_{\{\theta: g(\theta) \le g_0\}} \Pr[g_0 \notin \omega(\boldsymbol{X})|\theta]. \tag{9.1.20}$$

The confidence coefficient, on the other hand, is

$$1 - \sup_{\{\theta: g(\theta) = g_0\}} \Pr[g_0 \notin \omega(\boldsymbol{X})|\theta].$$

If we could prove that the supremum in Eq. (9.1.20) occurred at a θ for which $g(\theta) = g_0$, then the size of the test would be 1 minus the confidence coefficient. Most of the cases with which we shall deal in this book will have the property that the supremum in Eq. (9.1.20) does indeed occur at a θ for which $g(\theta) = g_0$. Example 9.1.16 is one such case. Example 9.1.13 is another. The following example is the general version of what we need in Example 9.1.13.

Example 9.1.17 One-Sided Tests and Confidence Intervals for a Normal Mean with Unknown Variance. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. Here $\theta = (\mu, \sigma^2)$. Let $g(\theta) = \mu$. In Theorem 8.5.1, we found that

$$\left(\overline{X}_n - T_{n-1}^{-1}(\gamma)\frac{\sigma'}{n^{1/2}}, \infty\right) \tag{9.1.21}$$

is a one-sided coefficient γ confidence interval for $g(\theta)$. Now, suppose that we use this interval to test hypotheses. We shall reject the null hypothesis that $\mu = \mu_0$ if μ_0 is not in the interval (9.1.21). It is easy to see that μ_0 is not in the interval (9.1.21) if and only if $\overline{X}_n \ge \mu_0 + \sigma' n^{-1/2} T_{n-1}^{-1}(\gamma)$. Such a test would seem to make sense for testing the hypotheses

$$H_0\colon \mu \le \mu_0, \qquad H_1\colon \mu > \mu_0. \tag{9.1.22}$$

In particular, in Example 9.1.13, the fact that 4 is not in the observed confidence interval means that the test constructed above (with $\mu_0 = 4$ and $\gamma = 0.9$) would reject $H_0 : \mu \le 4$ at level $\alpha_0 = 0.1$. ◀

The test constructed in Example 9.1.17 is another t test that we shall study in Sec. 9.5. In particular, we will show in Sec. 9.5 that this t test is a level $1 - \gamma$ test. In Exercise 19, you can find the one-sided confidence interval that corresponds to testing the reverse hypotheses.

Likelihood Ratio Tests

A very popular form of hypothesis test is the likelihood ratio test. We shall give a partial theoretical justification for likelihood ratio tests in Sec. 9.2. Such tests are based on the likelihood function $f_n(\boldsymbol{x}|\theta)$. (See Definition 7.2.3 on page 390.) The likelihood function tends to be highest near the true value of θ. Indeed, this is why maximum likelihood estimation works well in so many cases. Now, suppose that we wish to test the hypotheses

$$\begin{aligned} H_0\colon &\quad \theta \in \Omega_0, \\ H_1\colon &\quad \theta \in \Omega_1. \end{aligned} \tag{9.1.23}$$

In order to compare these two hypotheses, we might wish to see whether the likelihood function is higher on Ω_0 or on Ω_1, and if not, how much smaller the likelihood function is on Ω_0. When we computed M.L.E.'s, we maximized the likelihood function over the entire parameter space Ω. In particular, we calculated $\sup_{\theta\in\Omega} f_n(\boldsymbol{x}|\theta)$. If we restrict attention to H_0, then we can compute the largest value of the likelihood among those parameter values in Ω_0: $\sup_{\theta\in\Omega_0} f_n(\boldsymbol{x}|\theta)$. The ratio of these two suprema can then be used for testing the hypotheses (9.1.23).

Definition 9.1.11 Likelihood Ratio Test. The statistic

$$\Lambda(\boldsymbol{x}) = \frac{\sup_{\theta\in\Omega_0} f_n(\boldsymbol{x}|\theta)}{\sup_{\theta\in\Omega} f_n(\boldsymbol{x}|\theta)} \tag{9.1.24}$$

is called the *likelihood ratio statistic*. A *likelihood ratio test* of hypotheses (9.1.23) is to reject H_0 if $\Lambda(\boldsymbol{x}) \le k$ for some constant k.

In words, a likelihood ratio test rejects H_0 if the likelihood function on Ω_0 is sufficiently small compared to the likelihood function on all of Ω. Generally, k is chosen so that the test has a desired level α_0, if that is possible.

Example 9.1.18 Likelihood Ratio Test of Two-Sided Hypotheses about a Bernoulli Parameter. Suppose that we shall observe Y, the number of successes in n independent Bernoulli trials with unknown parameter θ. Consider the hypotheses $H_0 : \theta = \theta_0$ versus $H_0 : \theta \ne \theta_0$. After the value $Y = y$ has been observed, the likelihood function is

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

In this case, $\Omega_0 = \{\theta_0\}$ and $\Omega = [0, 1]$. The likelihood ratio statistic is

$$\Lambda(y) = \frac{\theta_0^y(1-\theta_0)^{n-y}}{\sup_{\theta\in[0,1]} \theta^y(1-\theta)^{n-y}}. \tag{9.1.25}$$

The supremum in the denominator of Eq. (9.1.25) can be found as in Example 7.5.4. The maximum occurs where θ equals the M.L.E., $\hat{\theta} = y/n$. So,

$$\Lambda(y) = \left(\frac{n\theta_0}{y}\right)^y \left(\frac{n(1-\theta_0)}{n-y}\right)^{n-y}.$$

It is not difficult to see that $\Lambda(y)$ is small for y near 0 and near n and largest near $y = n\theta_0$. As a specific example, suppose that $n = 10$ and $\theta_0 = 0.3$. Table 9.1 shows the 11 possible values of $\Lambda(y)$ for $y = 0, \ldots, 10$. If we desired a test with level of significance α_0, we would order the values of y according to values of $\Lambda(y)$ from smallest to largest and choose k so that the sum of the probabilities $\Pr(Y = y|\theta = 0.3)$ corresponding to those values of y with $\Lambda(y) \le k$ was at most α_0. For example, if $\alpha_0 = 0.05$, we see from Table 9.1 that we can add up the probabilities corresponding to $y = 10, 9, 8, 7, 0$ to get 0.039. But if we include $y = 6$, corresponding to the next smallest value of $\Lambda(y)$, the sum jumps to 0.076, which is too large. The set of $y \in \{10, 9, 8, 7, 0\}$ corresponds to $\Lambda(y) \le k$ for every k in the half-open interval [0.028, 0.147). The size of the test that rejects H_0 when $y \in \{10, 9, 8, 7, 0\}$ is 0.039. ◄

Likelihood Ratio Tests with Large Samples

Likelihood ratio tests are most popular in problems involving large sample sizes. The following result, whose precise statement and proof are beyond the scope of this text, shows how to use them in such cases.

Table 9.1 Values of the likelihood ratio statistic in Example 9.1.18

y	0	1	2	3	4	5	6	7	8	9	10
$\Lambda(y)$	0.028	0.312	0.773	1.000	0.797	0.418	0.147	0.034	0.005	3×10^{-4}	6×10^{-6}
$\Pr(Y = y \mid \theta = 0.3)$	0.028	0.121	0.233	0.267	0.200	0.103	0.037	0.009	0.001	1×10^{-4}	6×10^{-6}

Theorem 9.1.4 **Large-Sample Likelihood Ratio Tests.** Let Ω be an open subset of p-dimensional space, and suppose that H_0 specifies that k coordinates of θ are equal to k specific values. Assume that H_0 is true and that the likelihood function satisfies the conditions needed to prove that the M.L.E. is asymptotically normal and asymptotically efficient. (See page 523.) Then, as $n \to \infty$, $-2 \log \Lambda(\boldsymbol{X})$ converges in distribution to the χ^2 distribution with k degrees of freedom. ■

Example 9.1.19 **Likelihood Ratio Test of Two-Sided Hypotheses about a Bernoulli Parameter.** We shall apply the idea in Theorem 9.1.4 to the case at the end of Example 9.1.18. Set $\Omega = (0, 1)$ so that $p = 1$ and $k = 1$. To get an approximate level α_0 test, we would reject H_0 if $-2 \log \Lambda(y)$ is greater than the $1 - \alpha_0$ quantile of the χ^2 distribution with one degree of freedom. With $\alpha_0 = 0.05$, this quantile is 3.841. By taking logarithms of the numbers in the $\Lambda(y)$ row of Table 9.1, one sees that $-2 \log \Lambda(y) > 3.841$ for $y \in \{10, 9, 8, 7, 0\}$. Rejecting H_0 when $-2 \log \Lambda(y) > 3.841$ is then the same test as we constructed in Example 9.1.18. ◄

Theorem 9.1.4 can also be applied if the null hypothesis specifies that a collection of k functions of θ are equal to k specific values. For example, suppose that the parameter is $\theta = (\mu, \sigma^2)$, and we wish to test $H_0 : (\mu - 2)/\sigma = 1$ versus $H_1 : (\mu - 2)/\sigma \neq 1$. We could first transform to the equivalent parameter $\theta' = ([\mu - 2]/\sigma, \sigma)$ and then apply Theorem 9.1.4. Because of the invariance property of M.L.E.'s (Theorem 7.6.1, which extends to multidimensional parameters) one does not actually need to perform the transformation in order to compute Λ. One merely needs to maximize the likelihood function over the two sets Ω_0 and Ω and take the ratio.

On a final note, one must be careful not to apply Theorem 9.1.4 to problems of one-sided hypothesis testing. In such cases, the $\Lambda(\boldsymbol{X})$ usually has a distribution that is neither discrete nor continuous and doesn't converge to a χ^2 distribution. Also, Theorem 9.1.4 fails to apply when the parameter space Ω is a closed set and the null hypothesis is that θ takes a value on the boudary of Ω.

Hypothesis-Testing Terminology

We noted after Definition 9.1.1 that there is asymmetry in the terminology with regard to choosing between hypotheses. Both choices are stated relative to H_0, namely, to reject H_0 or not to reject H_0. When hypothesis testing was first being developed, there was controversy over whether alternative hypotheses should even be formulated. Focus centered on null hypotheses and whether or not to reject them. The operational meaning of "do not reject H_0" has never been articulated clearly. In particular, it does not mean that we should accept H_0 as true in any sense. Nor does it mean that we are necessarily more confident that H_0 is true than that it is false. For that matter, "reject H_0" does not mean that we are more confident that H_0 is false than that it is true.

Part of the problem is that hypothesis testing is set up as if it were a statistical decision problem, but neither a loss function nor a utility function is involved. Hence, we are not weighing the relative likelihoods of various hypotheses against the costs or benefits of making various decisions. In Sec. 9.8, we shall illustrate one method for treating the hypothesis-testing problem as a statistical decision problem. Many, but not all, of the popular testing procedures will turn out to have interpretations in the framework of decision problems. In the remainder of this chapter, we shall continue to develop the theory of hypothesis testing as it is generally practiced.

There are two other points of terminology that should be clarified here. The first concerns the terms "critical region" and "rejection region." Readers of other books might encounter either of the terms "critical region" or "rejection region" referring to either the set S_1 in Definition 9.1.4 or the set R in Definition 9.1.5. Those books generally define only one of the two terms. We choose to give the two sets S_1 and R different names because they are mathematically different objects. One, S_1, is a subset of the set of possible data vectors, while the other, R, is a subset of the set of possible values of a test statistic. Each has its use in different parts of the development of hypothesis testing. In most practical problems, tests are more easily expressed in terms of test statistics and rejection regions. For proving some theorems in Sec. 9.2, it is more convenient to define tests in terms of critical regions.

The final point of terminology concerns the terms "level of significance" and "size," as well as the term "level α_0 test." Some authors define level of significance (or significance level) for a test using a phrase such as "the probability of type I error" or "the probability that the data lie in the critical region when the null hypothesis is true." If the null hypothesis is simple, these phrases are easily understood, and they match what we defined as the size of the test in such cases. On the other hand, if the null hypothesis is composite, such phrases are ill-defined. For each $\theta \in \Omega_0$, there will usually be a different probability that the test rejects H_0. Which, if any, is the level of significance? We have defined the size of a test to be the supremum of all of these probabilities. We have said that the test "has level of significance α_0" if the size is less than or equal to α_0. This means that a test has one size but many levels of significance. Every number from the size up to 1 is a level of significance. There is a sound reason for distinguishing the concepts of size and level of significance. In Example 9.1.9, the investigator wants to constrain the probability of type I error to be less than 0.1. The test statistic Y has a discrete distribution, and we saw that no test with size 0.1 is available. In that example, the investigator needed to choose a test whose size was 0.0473. This test still has level of significance 0.1 and is a level 0.1 test, despite having a different size. There are other more complicated situations in which one can construct a test δ that satisfies Eq. (9.1.6), that is, it has level of significance α_0, but for which it is not possible (without sophisticated numerical methods) to compute the actual size. An investigator who insists on using a particular level of significance α_0 can use such a test, and call it a level α_0 test, without being able to compute its size exactly. The most common example of this latter situation is one in which we wish to test hypotheses concerning two parameters simultaneously. For example, let $\theta = (\theta_1, \theta_2)$, and suppose that we wish to test the hypotheses

$$H_0: \theta_1 = 0 \text{ and } \theta_2 = 1 \quad \text{versus} \quad H_1: \theta_1 \neq 0 \text{ or } \theta_2 \neq 1 \text{ or both.} \qquad (9.1.26)$$

The following result gives a way to contruct a level α_0 test of H_0.

Theorem 9.1.5 For $i = 1, \ldots, n$, let $H_{0,i}$ be a null hypothesis, and let δ_i be a level $\alpha_{0,i}$ test of $H_{0,i}$. Define the combined null hypothesis H_0 that all of $H_{0,1}, \ldots, H_{0,n}$ are simultaneously true. Let δ be the test that rejects H_0 if at least one of $\delta_1, \ldots, \delta_n$ rejects its corresponding null hypothesis. Then δ is a level $\sum_{i=1}^{n} \alpha_{0,i}$ test of H_0.

Proof For $i = 1, \ldots, n$, let A_i be the event that δ_i rejects $H_{0,i}$. Apply Theorem 1.5.8. ■

To test H_0 in (9.1.26), find two tests δ_1 and δ_2 such that δ_1 is a test with size $\alpha_0/2$ for testing $\theta_1 = 0$ versus $\theta_1 \neq 0$ and δ_2 is a test with size $\alpha_0/2$ for testing $\theta_2 = 1$ versus $\theta_2 \neq 1$. Let δ be the test that rejects H_0 if either δ_1 rejects $\theta_1 = 0$ or δ_2 rejects $\theta_2 = 1$ or both. Theorem 9.1.5 says that δ is a level α_0 test of H_0 versus H_1, but its exact size requires us to be able to calculate the probability that both δ_1 and δ_2 simultaneously reject their corresponding null hypotheses. Such a calculation is often intractable.

Finally, our definition of level of significance matches nicely with the use of p-values, as pointed out immediately after Definition 9.1.9.

Summary

Hypothesis testing is the problem of deciding whether θ lies in a particular subset Ω_0 of the parameter space or in its complement Ω_1. The statement that $\theta \in \Omega_0$ is called the null hypothesis and is denoted by H_0. The alternative hypothesis is the statement $H_1 : \theta \in \Omega_1$. If S is the set of all possible data values (vectors) that we might observe, a subset $S_1 \subset S$ is called the critical region of a test of H_0 versus H_1 if we choose to reject H_0 whenever the observed data $\boldsymbol{X}$ are in S_1 and not reject H_0 whenever $\boldsymbol{X} \notin S_1$. The power function of this test δ is $\pi(\theta|\delta) = \Pr(\boldsymbol{X} \in S_1|\theta)$. The size of the test δ is $\sup_{\theta \in \Omega_0} \pi(\theta|\delta)$. A test is said to be a level α_0 test if its size is at most α_0. The null hypothesis H_0 is simple if Ω_0 is a set with only one point; otherwise, H_0 is composite. Similarly, H_1 is simple if Ω_1 has a single point, and H_1 is composite otherwise. A type I error is rejecting H_0 when it is true. A type II error is not rejecting H_0 when it is false.

Hypothesis tests are typically constructed by using a test statistic T. The null hypothesis is rejected if T lies in some interval or if T lies outside of some interval. The interval is chosen to make the test have a desired significance level. The p-value is a more informative way to report the results of a test. The p-value can be computed easily whenever our test has the form "reject H_0 if $T \geq c$" for some statistic T. The p-value when $T = t$ is observed equals $\sup_{\theta \in \Omega_0} \Pr(T \geq t|\theta)$. We also showed how a confidence set can be considered as a way of reporting the results of a test of hypotheses. A coefficient $1 - \alpha_0$ confidence set for θ is the set of all $\theta_0 \in \Omega$, such that we would not reject $H_0 : \theta = \theta_0$ using a level α_0 test. These confidence sets are intervals when we test hypotheses about a one-dimensional parameter or a one-dimensional function of the parameter.

Exercises

1. Let X have the exponential distribution with parameter β. Suppose that we wish to test the hypotheses $H_0 : \beta \ge 1$ versus $H_1 : \beta < 1$. Consider the test procedure δ that rejects H_0 if $X \ge 1$.

a. Determine the power function of the test.

b. Compute the size of the test.

2. Suppose that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, and that the following hypotheses are to be tested:

$$H_0: \quad \theta \ge 2,$$
$$H_1: \quad \theta < 2.$$

Let $Y_n = \max\{X_1, \dots, X_n\}$, and consider a test procedure such that the critical region contains all the outcomes for which $Y_n \le 1.5$.

a. Determine the power function of the test.

b. Determine the size of the test.

3. Suppose that the proportion p of defective items in a large population of items is unknown, and that it is desired to test the following hypotheses:

$$H_0: \quad p = 0.2,$$
$$H_1: \quad p \ne 0.2.$$

Suppose also that a random sample of 20 items is drawn from the population. Let Y denote the number of defective items in the sample, and consider a test procedure δ such that the critical region contains all the outcomes for which either $Y \ge 7$ or $Y \le 1$.

a. Determine the value of the power function $\pi(p|\delta)$ at the points $p = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$, and 1; sketch the power function.

b. Determine the size of the test.

4. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance 1. Suppose also that μ_0 is a certain specified number, and that the following hypotheses are to be tested:

$$H_0: \quad \mu = \mu_0,$$
$$H_1: \quad \mu \ne \mu_0.$$

Finally, suppose that the sample size n is 25, and consider a test procedure such that H_0 is to be rejected if $|\overline{X}_n - \mu_0| \ge c$. Determine the value of c such that the size of the test will be 0.05.

5. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Classify each of the following hypotheses as either simple or composite:

a. H_0: $\mu = 0$ and $\sigma = 1$

b. H_0: $\mu > 3$ and $\sigma < 1$

c. H_0: $\mu = -2$ and $\sigma^2 < 5$

d. H_0: $\mu = 0$

6. Suppose that a single observation X is to be taken from the uniform distribution on the interval $\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, and suppose that the following hypotheses are to be tested:

$$H_0: \quad \theta \le 3,$$
$$H_1: \quad \theta \ge 4.$$

Construct a test procedure δ for which the power function has the following values: $\pi(\theta|\delta) = 0$ for $\theta \le 3$ and $\pi(\theta|\delta) = 1$ for $\theta \ge 4$.

7. Return to the situation described in Example 9.1.7. Consider a different test δ^* that rejects H_0 if $Y_n \le 2.9$ or $Y_n \ge 4.5$. Let δ be the test described in Example 9.1.7.

a. Prove that $\pi(\theta|\delta^*) = \pi(\theta|\delta)$ for all $\theta \le 4$.

b. Prove that $\pi(\theta|\delta^*) < \pi(\theta|\delta)$ for all $\theta > 4$.

c. Which of the two tests seems better for testing the hypotheses (9.1.8)?

8. Assume that $X_1, \dots, X_n$ are i.i.d. with the normal distribution that has mean μ and variance 1. Suppose that we wish to test the hypotheses

$$H_0: \quad \mu \le \mu_0,$$
$$H_1: \quad \mu > \mu_0.$$

Consider the test that rejects H_0 if $Z \ge c$, where Z is defined in Eq. (9.1.10).

a. Show that $\Pr(Z \ge c|\mu)$ is an increasing function of μ.

b. Find c to make the test have size α_0.

9. Assume that $X_1, \dots, X_n$ are i.i.d. with the normal distribution that has mean μ and variance 1. Suppose that we wish to test the hypotheses

$$H_0: \quad \mu \ge \mu_0,$$
$$H_1: \quad \mu < \mu_0.$$

Find a test statistic T such that, for every c, the test δ_c that rejects H_0 when $T \ge c$ has power function $\pi(\mu|\delta_c)$ that is decreasing in μ.

10. In Exercise 8, assume that $Z = z$ is observed. Find a formula for the p-value.

11. Assume that $X_1, \dots, X_9$ are i.i.d. having the Bernoulli distribution with parameter p. Suppose that we wish to test the hypotheses

$$H_0: \quad p = 0.4,$$
$$H_1: \quad p \ne 0.4.$$

Let $Y = \sum_{i=1}^{9} X_i$.

a. Find c_1 and c_2 such that

$$\Pr(Y \le c_1 | p = 0.4) + \Pr(Y \ge c_2 | p = 0.4)$$

is as close as possible to 0.1 without being larger than 0.1.

b. Let δ be the test that rejects H_0 if either $Y \le c_1$ or $Y \ge c_2$. What is the size of the test δ_c?

c. Draw a graph of the power function of δ_c.

12. Consider a single observation X from a Cauchy distribution centered at θ. That is, the p.d.f. of X is

$$f(x|\theta) = \frac{1}{\pi[1 + (x - \theta)^2]}, \quad \text{for } -\infty < x < \infty.$$

Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \theta \le \theta_0, \\ H_1&: \quad \theta > \theta_0. \end{aligned}$$

Let δ_c be the test that rejects H_0 if $X \ge c$.

a. Show that $\pi(\theta|\delta_c)$ is an increasing function of θ.

b. Find c to make δ_c have size 0.05.

c. If $X = x$ is observed, find a formula for the p-value.

13. Let X have the Poisson distribution with mean θ. Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \theta \le 1.0, \\ H_1&: \quad \theta > 1.0. \end{aligned}$$

Let δ_c be the test that rejects H_0 if $X \ge c$. Find c to make the size of δ_c as close as possible to 0.1 without being larger than 0.1.

14. Let $X_1, \ldots, X_n$ be i.i.d. with the exponential distribution with parameter θ. Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \theta \ge \theta_0, \\ H_1&: \quad \theta < \theta_0. \end{aligned}$$

Let $X = \sum_{i=1}^{n} X_i$. Let δ_c be the test that rejects H_0 if $X \ge c$.

a. Show that $\pi(\theta|\delta_c)$ is a decreasing function of θ.

b. Find c in order to make δ_c have size α_0.

c. Let $\theta_0 = 2$, $n = 1$, and $\alpha_0 = 0.1$. Find the precise form of the test δ_c and sketch its power function.

15. Let X have the uniform distribution on the interval $[0, \theta]$, and suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \theta \le 1, \\ H_1&: \quad \theta > 1. \end{aligned}$$

We shall consider test procedures of the form "reject H_0 if $X \ge c$." For each possible value x of X, find the p-value if $X = x$ is observed.

16. Consider the confidence interval found in Exercise 5 in Sec. 8.5. Find the collection of hypothesis tests that are equivalent to this interval. That is, for each $c > 0$, find a test δ_c of the null hypothesis $H_{0,c} : \sigma^2 = c$ versus some alternative such that δ_c rejects $H_{0,c}$ if and only if c is not in the interval. Write the test in terms of a test statistic $T = r(\boldsymbol{X})$ being in or out of some nonrandom interval that depends on c.

17. Let $X_1, \ldots, X_n$ be i.i.d. with a Bernoulli distribution that has parameter p. Let $Y = \sum_{i=1}^{n} X_i$. We wish to find a coefficient γ confidence interval for p of the form $(q(y), 1)$. Prove that, if $Y = y$ is observed, then $q(y)$ should be chosen to be the smallest value p_0 such that $\Pr(Y \ge y | p = p_0) \ge 1 - \gamma$.

18. Consider the situation described immediately before Eq. (9.1.12). Prove that the expression (9.1.12) equals the smallest α_0 such that we would reject H_0 at level of significance α_0.

19. Return to the situation described in Example 9.1.17. Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \mu \ge \mu_0, \\ H_1&: \quad \mu < \mu_0 \end{aligned} \tag{9.1.27}$$

at level α_0. It makes sense to reject H_0 if $\overline{X}_n$ is small. Construct a one-sided coefficient $1 - \alpha_0$ confidence interval for μ such that we can reject H_0 if μ_0 is not in the interval. Make sure that the test formed in this way rejects H_0 if $\overline{X}_n$ is small.

20. Prove Theorem 9.1.3.

21. Return to the situations described in Example 9.1.17 and Exercise 19. We wish to compare what might happen if we switch the null and alternative hypotheses. That is, we want to compare the results of testing the hypotheses in (9.1.22) at level α_0 to the results of testing the hypotheses in (9.1.27) at level α_0.

a. Let $\alpha_0 < 0.5$. Prove that there are no possible data sets such that we would reject both of the null hypotheses simultaneously. That is, for every possible $\overline{X}_n$ and σ', we must fail to reject at least one of the two null hypotheses.

b. Let $\alpha_0 < 0.5$. Prove that there are data sets that would lead to failing to reject both null hypotheses. Also prove that there are data sets that would lead to rejecting each of the null hypotheses while failing to reject the other.

c. Let $\alpha_0 > 0.5$. Prove that there are data sets that would lead to rejecting both null hypotheses.

★ 9.2 Testing Simple Hypotheses

The simplest hypothesis-testing situation is that in which there are only two possible values of the parameter. In such cases, it is possible to identify a collection of test procedures that have certain optimal properties.

Introduction

Example 9.2.1 Service Times in a Queue. In Example 3.7.5, we modeled the service times $\boldsymbol{X} = (X_1, \dots, X_n)$ of n customers in a queue as having the joint distribution with joint p.d.f.

$$f_1(\boldsymbol{x}) = \begin{cases} \dfrac{2(n!)}{\left(2 + \sum_{i=1}^n x_i\right)^{n+1}} & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9.2.1}$$

Suppose that a service manager is not sure how well this joint distribution describes the service times. As an alternative, she proposes to model the service times as a random sample of exponential random variables with parameter 1/2. This model says that the joint p.d.f. is

$$f_0(\boldsymbol{x}) = \begin{cases} \dfrac{1}{2^n} \exp\left(-\dfrac{1}{2}\displaystyle\sum_{i=1}^n x_i\right) & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9.2.2}$$

For illustration, Fig. 9.5 shows both of these p.d.f.'s for the case of $n = 1$. If the manager observes several service times, how can she test which of the two distributions appears to describe the data? ◀

In this section, we shall consider problems of testing hypotheses in which a vector of observations comes from one of two possible joint distributions, and the statistician must decide from which distribution the vector actually came. In many problems, each of the two joint distributions is actually the distribution of a random sample from a univariate distribution. However, nothing that we present in this section will depend on whether or not the observations form a random sample. In Example 9.2.1, one of the joint distributions is that of a random sample, but the other is not. In problems of this type, the parameter space Ω contains exactly two points, and both the null hypothesis and the alternative hypothesis are simple.

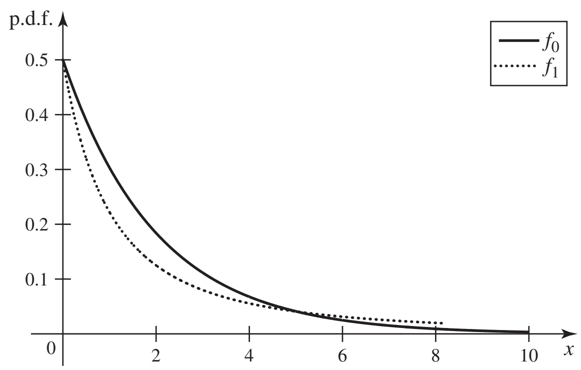


Figure 9.5 Graphs of the two competing p.d.f.'s in Example 9.2.1 with $n = 1$.

Specifically, we shall assume that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ comes from a distribution for which the joint p.d.f., p.f., or p.f./p.d.f. is either $f_0(\boldsymbol{x})$ or $f_1(\boldsymbol{x})$. To correspond with notation earlier and later in the book, we can introduce a parameter space $\Omega = \{\theta_0, \theta_1\}$ and let $\theta = \theta_i$ stand for the case in which the data have p.d.f., p.f., or p.f./p.d.f. $f_i(\boldsymbol{x})$ for $i = 0, 1$. We are then interested in testing the following simple hypotheses:

$$\begin{aligned} H_0&: \quad \theta = \theta_0, \\ H_1&: \quad \theta = \theta_1. \end{aligned} \tag{9.2.3}$$

In this case, $\Omega_0 = \{\theta_0\}$ and $\Omega_1 = \{\theta_1\}$ are both singleton sets.

For the special case in which $\boldsymbol{X}$ is a random sample from a distribution with univariate p.d.f. or p.f. $f(x|\theta)$, we then have, for $i = 0$ or $i = 1$,

$$f_i(\boldsymbol{x}) = f(x_1|\theta_i) f(x_2|\theta_i) \cdots f(x_n|\theta_i).$$

The Two Types of Errors

When a test of the hypotheses (9.2.3) is being carried out, we have special notation for the probabilities of type I and type II errors. For each test procedure δ, we shall let $\alpha(\delta)$ denote the probability of an error of type I and shall let $\beta(\delta)$ denote the probability of an error of type II. Thus,

$$\begin{aligned} \alpha(\delta) &= \Pr(\text{Rejecting } H_0 | \theta = \theta_0), \\ \beta(\delta) &= \Pr(\text{Not Rejecting } H_0 | \theta = \theta_1). \end{aligned}$$

Example 9.2.2 Service Times in a Queue. The manager in Example 9.2.1 looks at the two p.d.f.'s in Fig. 9.5 and decides that f_1 gives higher probability to large service times than does f_0. So she decides to reject $H_0 : \theta = \theta_0$ if the service times are large. Specifically, suppose that she observes $n = 1$ service time, X_1. The test δ that she chooses rejects H_0 if $X_1 \geq 4$. The two error probabilities can be calculated from the two different possible distributions of X_1. Given $\theta = \theta_0$, X_1 has the exponential distribution with parameter 0.5. The c.d.f. of this distribution is $F_0(x) = 1 - \exp(-0.5x)$ for $x \geq 0$. The type I error probability is the probability that $X_1 \geq 4$, which equals $\alpha(\delta) = 0.135$. Given $\theta = \theta_1$, the distribution of X_1 has the p.d.f. $2/(2 + x_1)^2$ for $x_1 \geq 0$. The c.d.f. is then $F_1(x) = 1 - 2/(2 + x)$, for $x \geq 0$. The type II error probability is $\beta(\delta) = \Pr(X_1 < 4) = F_1(4) = 0.667$. ◀

It is desirable to find a test procedure for which the probabilities $\alpha(\delta)$ and $\beta(\delta)$ of the two types of error will be small. For a given sample size, it is typically not possible to find a test procedure for which both $\alpha(\delta)$ and $\beta(\delta)$ will be arbitrarily small. Therefore, we shall now show how to construct a procedure for which the value of a specific linear combination of α and β will be minimized.

Optimal Tests

Minimizing a Linear Combination Suppose that a and b are specified positive constants, and it is desired to find a procedure δ for which $a\alpha(\delta) + b\beta(\delta)$ will be a minimum. Theorem 9.2.1 shows that a procedure that is optimal in this sense has a very simple form. In Sec. 9.8, we shall give a rationale for choosing a test to minimize a linear combination of the error probabilities.

Theorem 9.2.1 Let δ^* denote a test procedure such that the hypothesis H_0 is not rejected if $af_0(\boldsymbol{x}) > bf_1(\boldsymbol{x})$ and the hypothesis H_0 is rejected if $af_0(\boldsymbol{x}) < bf_1(\boldsymbol{x})$. The null hypothesis H_0 can be either rejected or not if $af_0(\boldsymbol{x}) = bf_1(\boldsymbol{x})$. Then for every other test procedure δ,

$$a\alpha(\delta^*) + b\beta(\delta^*) \le a\alpha(\delta) + b\beta(\delta). \tag{9.2.4}$$

Proof For convenience, we shall present the proof for a problem in which the random sample $X_1, \ldots, X_n$ is drawn from a discrete distribution. In this case, $f_i(\boldsymbol{x})$ represents the joint p.f. of the observations in the sample when H_i is true ($i = 0, 1$). If the sample comes from a continuous distribution, in which case $f_i(\boldsymbol{x})$ is a joint p.d.f., then each of the sums that will appear in this proof should be replaced by an n-dimensional integral.

If we let S_1 denote the critical region of an arbitrary test procedure δ, then S_1 contains every sample outcome $\boldsymbol{x}$ for which δ specifies that H_0 should be rejected, and $S_0 = S_1^c$ contains every outcome $\boldsymbol{x}$ for which H_0 should not be rejected. Therefore,

$$\begin{aligned} a\alpha(\delta) + b\beta(\delta) &= a \sum_{\boldsymbol{x}\in S_1} f_0(\boldsymbol{x}) + b \sum_{\boldsymbol{x}\in S_0} f_1(\boldsymbol{x}) \\ &= a \sum_{\boldsymbol{x}\in S_1} f_0(\boldsymbol{x}) + b \left[1 - \sum_{\boldsymbol{x}\in S_1} f_1(\boldsymbol{x}) \right] \\ &= b + \sum_{\boldsymbol{x}\in S_1} [af_0(\boldsymbol{x}) - bf_1(\boldsymbol{x})]. \end{aligned} \tag{9.2.5}$$

It follows from Eq. (9.2.5) that the value of the linear combination $a\alpha(\delta) + b\beta(\delta)$ will be a minimum if the critical region S_1 is chosen so that the value of the final summation in Eq. (9.2.5) is a minimum. Furthermore, the value of this summation will be a minimum if the summation includes every point $\boldsymbol{x}$ for which $af_0(\boldsymbol{x}) - bf_1(\boldsymbol{x}) < 0$ and includes no point $\boldsymbol{x}$ for which $af_0(\boldsymbol{x}) - bf_1(\boldsymbol{x}) > 0$. In other words, $a\alpha(\delta) + b\beta(\delta)$ will be a minimum if the critical region S_1 is chosen to include every point $\boldsymbol{x}$ such that $af_0(\boldsymbol{x}) < bf_1(\boldsymbol{x})$ and exclude every point $\boldsymbol{x}$ such that this inequality is reversed. If $af_0(\boldsymbol{x}) = bf_1(\boldsymbol{x})$ for some point $\boldsymbol{x}$, then it is irrelevant whether or not $\boldsymbol{x}$ is included in S_1, because the corresponding term would contribute zero to the final summation in Eq. (9.2.5). The critical region described above corresponds to the test procedure δ^* defined in the statement of the theorem. ■

The ratio $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ is sometimes called the *likelihood ratio* of the sample. It is related to, but not the same as, the likelihood ratio statistic from Definition 9.1.11. In the present context, the likelihood ratio statistic $\Lambda(\boldsymbol{x})$ would equal $f_0(\boldsymbol{x})/ \max\{f_0(\boldsymbol{x}), f_1(\boldsymbol{x})\}$. In particular, the likelihood ratio $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ is large when $\Lambda(\boldsymbol{x})$ is small, and vice versa. In fact,

$$\Lambda(\boldsymbol{x}) = \begin{cases} \left(\dfrac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} \right)^{-1} & \text{if } f_0(\boldsymbol{x}) \le f_1(\boldsymbol{x}) \\ 1 & \text{otherwise.} \end{cases}$$

The important point to remember about this confusing choice of names is the following: The theoretical justification for tests based on the likelihood ratio defined here (provided in Theorems 9.2.1 and 9.2.2) is the rationale for expecting the likelihood ratio tests of Defintion 9.1.11 to be sensible.

When $a, b > 0$, Theorem 9.2.1 can be reworded as follows.

Corollary 9.2.1 Assume the conditions of Theorem 9.2.1, and assume that $a > 0$ and $b > 0$. Then the test δ for which the value of $a\alpha(\delta) + b\beta(\delta)$ is a minimum rejects H_0 when the likelihood ratio exceeds a/b and does not reject H_0 when the likelihood ratio is less than a/b. ■

Example 9.2.3 Service Times in a Queue. Instead of rejecting H_0 if $X_1 \geq 4$ in Example 9.2.2, the manager could apply Theorem 9.2.1. She must choose two numbers a and b to balance the two types of error. Suppose that she chooses them to be equal to each other. Then the test will be to reject H_0 if $f_1(x_1)/f_0(x_1) > 1$. That is, if

$$\frac{4}{(2+x_1)^2}\exp\left(\frac{x_1}{2}\right) > 1. \tag{9.2.6}$$

At $x_1 = 0$ the left side of Eq. (9.2.6) equals 1, and it decreases until $x_1 = 2$ and then increases ever after. Hence, Eq. (9.2.6) holds for all values of $x_1 > c$ where c is the unique strictly positive value where the left side of Eq. (9.2.6) equals 1. By numerical approximation, we find that this value is $x_1 = 5.025725$. The type I and type II error probabilities for the test δ^* that rejects H_0 if $X_1 > 5.025725$ are

$$\alpha(\delta^*) = 1 - F_0(5.025725) = \exp(-2.513) = 0.081,$$

$$\beta(\delta^*) = F_1(5.025725) = 1 - \frac{2}{7.026} = 0.715.$$

The sum of these error probabilities is 0.796. By comparison, the sum of the two error probabilities in Example 9.2.2 is 0.802, slightly higher. ◀

Minimizing the Probability of an Error of Type II Next, suppose that the probability $\alpha(\delta)$ of an error of type I is not permitted to be greater than a specified level of significance, and it is desired to find a procedure δ for which $\beta(\delta)$ will be a minimum. In this problem, we can apply the following result, which is closely related to Theorem 9.2.1 and is known as the *Nayman-Pearson lemma* in honor of the statisticians J. Neyman and E. S. Pearson, who developed these ideas in 1933.

Theorem 9.2.2 Nayman-Pearson lemma. Suppose that δ' is a test procedure that has the following form for some constant $k > 0$: The hypothesis H_0 is not rejected if $f_1(\boldsymbol{x}) < kf_0(\boldsymbol{x})$ and the hypothesis H_0 is rejected if $f_1(\boldsymbol{x}) > kf_0(\boldsymbol{x})$. The null hypothesis H_0 can be either rejected or not if $f_1(\boldsymbol{x}) = kf_0(\boldsymbol{x})$. If δ is another test procedure such that $\alpha(\delta) \leq \alpha(\delta')$, then it follows that $\beta(\delta) \geq \beta(\delta')$. Furthermore, if $\alpha(\delta) < \alpha(\delta')$, then $\beta(\delta) > \beta(\delta')$.

Proof From the description of the procedure δ' and from Theorem 9.2.1, it follows that for every test procedure δ,

$$k\alpha(\delta') + \beta(\delta') \leq k\alpha(\delta) + \beta(\delta). \tag{9.2.7}$$

If $\alpha(\delta) \leq \alpha(\delta')$, then it follows from the relation (9.2.7) that $\beta(\delta) \geq \beta(\delta')$. Also, if $\alpha(\delta) < \alpha(\delta')$, then it follows that $\beta(\delta) > \beta(\delta')$. ■

To illustrate the use of the Nayman-Pearson lemma, we shall suppose that a statistician wishes to use a test procedure for which $\alpha(\delta) = \alpha_0$ and $\beta(\delta)$ is a minimum. According to the lemma, she should try to find a value of k for which $\alpha(\delta') = \alpha_0$. The procedure δ' will then have the minimum possible value of $\beta(\delta)$. If the distribution from which the sample is taken is continuous, then it is usually (but not always) possible to find a value of k such that $\alpha(\delta')$ is equal to a specified value such as α_0. However, if the distribution from which the sample is taken is discrete, then it is typically not possible to choose k so that $\alpha(\delta')$ is equal to a specified value. These remarks are considered further in the following examples and in the exercises at the end of this section.

Example 9.2.4

Service Times in a Queue. In Example 9.2.3, the distribution of X_1 is continuous, and we can find a value k such that the test δ' that results from Theorem 9.2.2 has $\alpha(\delta') = 0.07$, say. The test δ^* in Example 9.2.3 has $\alpha(\delta^*) > 0.07$ and $k = 1$. We will need a larger value of k in order to get the type I error probability down to 0.07. As we noted in Example 9.2.3, the left side of Eq. (9.2.6) is increasing for $x_1 > 2$, and hence the set of x_1 values such that

$$\frac{4}{(2+x_1)^2}\exp\left(\frac{x_1}{2}\right) > k \tag{9.2.8}$$

will be an interval of the form (c, ∞) where c is the unique value that makes the left side of Eq. (9.2.8) equal to k. The resulting test will then have the form "reject H_0 if $X_1 \geq c$." At this point, we don't care any more about k because we just need to choose c to make sure that $\Pr(X_1 \geq c|\theta = \theta_0) = 0.07$. That is, we need $1 - F_0(c) = 0.07$. Recall that $F_0(c) = 1 - \exp(-0.5c)$, so $c = -2\log(0.07) = 5.318$. We can then compute $\beta(\delta') = F_1(5.318) = 0.727$. This test is very close to δ^* from Example 9.2.3. ◀

Example 9.2.5

Random Sample from a Normal Distribution. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random sample from the normal distribution with unknown mean θ and known variance 1, and the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \theta = 0, \\ H_1&: \quad \theta = 1. \end{aligned} \tag{9.2.9}$$

We shall begin by determining a test procedure for which $\beta(\delta)$ will be a minimum among all test procedures for which $\alpha(\delta) \leq 0.05$.

When H_0 is true, the variables $X_1, \ldots, X_n$ form a random sample from the standard normal distribution. When H_1 is true, these variables form a random sample from the normal distribution for which both the mean and the variance are 1. Therefore,

$$f_0(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \tag{9.2.10}$$

and

$$f_1(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - 1)^2\right]. \tag{9.2.11}$$

After some algebraic simplification, the likelihood ratio $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ can be written in the form

$$\frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} = \exp\left[n\left(\overline{x}_n - \frac{1}{2}\right)\right]. \tag{9.2.12}$$

It now follows from Eq. (9.2.12) that rejecting the hypothesis H_0 when the likelihood ratio is greater than a specified positive constant k is equivalent to rejecting H_0 when the sample mean $\overline{x}_n$ is greater than $(1/2) + (1/n)\log k$.

Let $k' = (1/2) + (1/n)\log k$, and suppose that we can find a value of k' such that

$$\Pr\left(\overline{X}_n > k'|\theta = 0\right) = 0.05. \tag{9.2.13}$$

Then the procedure δ', which rejects H_0 when $\overline{X}_n > k'$, will satisfy $\alpha(\delta') = 0.05$. Furthermore, by the Nayman-Pearson lemma, δ' will be an optimal procedure in the sense of minimizing the value of $\beta(\delta)$ among all procedures for which $\alpha(\delta) \leq 0.05$.

It is easy to see that the value of k' that satisfies Eq. (9.2.13) must be the 0.95 quantile of the distribution of $\overline{X}_n$ given $\theta = 0$. When $\theta = 0$, the distribution of $\overline{X}_n$ is the normal distribution with mean 0 and variance $1/n$. Therefore, its 0.95 quantile is $0 + \Phi^{-1}(0.95)n^{-1/2}$, where Φ^{-1} is the standard normal quantile function. From a table of the standard normal distribution, it is found that the 0.95 quantile of the standard normal distribution is 1.645, so $k' = 1.645n^{-1/2}$.

In summary, among all test procedures for which $\alpha(\delta) \le 0.05$, the procedure that rejects H_0 when $\overline{X}_n > 1.645n^{-1/2}$ has the smallest probability of type II error.

Next, we shall determine the probability $\beta(\delta')$ of an error of type II for this procedure δ'. Since $\beta(\delta')$ is the probability of not rejecting H_0 when H_1 is true,

$$\beta(\delta') = \Pr(\overline{X}_n < 1.645n^{-1/2} | \theta = 1). \tag{9.2.14}$$

When $\theta = 1$, the distribution of $\overline{X}_n$ is the normal distribution with mean 1 and variance $1/n$. The probability in Eq. (9.2.14) can then be written as

$$\beta(\delta') = \Phi\left(\frac{1.645n^{-1/2} - 1}{n^{-1/2}}\right) = \Phi(1.645 - n^{1/2}). \tag{9.2.15}$$

For instance, when $n = 9$, it is found from a table of the standard normal distribution that

$$\beta(\delta') = \Phi(-1.355) = 1 - \Phi(1.355) = 0.0877.$$

Finally, for this same random sample and the same hypotheses (9.2.9), we shall determine the test procedure δ_0 for which the value of $2\alpha(\delta) + \beta(\delta)$ is a minimum, and we shall calculate the value of $2\alpha(\delta_0) + \beta(\delta_0)$ when $n = 9$.

It follows from Theorem 9.2.1 that the procedure δ_0 for which $2\alpha(\delta) + \beta(\delta)$ is a minimum rejects H_0 when the likelihood ratio is greater than 2. By Eq. (9.2.12), this procedure is equivalent to rejecting H_0 when $\overline{X}_n > (1/2) + (1/n) \log 2$. Thus, when $n = 9$, the optimal procedure δ_0 rejects H_0 when $\overline{X}_n > 0.577$. For this procedure we then have

$$\alpha(\delta_0) = \Pr(\overline{X}_n > 0.577 | \theta = 0) \tag{9.2.16}$$

and

$$\beta(\delta_0) = \Pr(\overline{X}_n < 0.577 | \theta = 1). \tag{9.2.17}$$

Since $\overline{X}_n$ has the normal distribution with mean θ and variance $1/n$, we have

$$\alpha(\delta_0) = 1 - \Phi\left(\frac{0.577 - 0}{1/3}\right) = 1 - \Phi(1.731) = 0.0417$$

and

$$\beta(\delta_0) = \Phi\left(\frac{0.577 - 1}{1/3}\right) = \Phi(-1.269) = 0.1022.$$

The minimum value of $2\alpha(\delta) + \beta(\delta)$ is therefore

$$2\alpha(\delta_0) + \beta(\delta_0) = 2(0.0417) + (0.1022) = 0.1856. \qquad \blacktriangleleft$$

Example 9.2.6

Sampling from a Bernoulli Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p, and the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad p = 0.2, \\ H_1&: \quad p = 0.4. \end{aligned} \tag{9.2.18}$$

It is desired to find a test procedure for which $\alpha(\delta) = 0.05$ and $\beta(\delta)$ is a minimum.

In this example, each observed value x_i must be either 0 or 1. If we let $y = \sum_{i=1}^{n} x_i$, then the joint p.f. of $X_1, \ldots, X_n$ when $p = 0.2$ is

$$f_0(\boldsymbol{x}) = (0.2)^y (0.8)^{n-y} \tag{9.2.19}$$

and the joint p.f. when $p = 0.4$ is

$$f_1(\boldsymbol{x}) = (0.4)^y (0.6)^{n-y}. \tag{9.2.20}$$

Hence, the likelihood ratio is

$$\frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} = \left(\frac{3}{4}\right)^n \left(\frac{8}{3}\right)^y. \tag{9.2.21}$$

It follows that rejecting H_0 when the likelihood ratio is greater than a specified positive constant k is equivalent to rejecting H_0 when y is greater than k', where

$$k' = \frac{\log k + n \log(4/3)}{\log(8/3)}. \tag{9.2.22}$$

To find a test procedure for which $\alpha(\delta) = 0.05$ and $\beta(\delta)$ is a minimum, we use the Nayman-Pearson lemma. If we let $Y = \sum_{i=1}^{n} X_i$, we should try to find a value of k' such that

$$\Pr(Y > k' | p = 0.2) = 0.05. \tag{9.2.23}$$

When the hypothesis H_0 is true, the random variable Y has the binomial distribution with parameters n and $p = 0.2$. However, because of the discreteness of this distribution, it generally will not be possible to find a value of k' for which Eq. (9.2.23) is satisfied. For example, suppose that $n = 10$. Then it is found from a table of the binomial distribution that $\Pr(Y > 4 | p = 0.2) = 0.0328$ and also $\Pr(Y > 3 | p = 0.2) = 0.1209$. Therefore, there is no critical region of the desired form for which $\alpha(\delta) = 0.05$. If it is desired to use a level 0.05 test δ based on the likelihood ratio as specified by the Nayman-Pearson lemma, then one must reject H_0 when $Y > 4$ and $\alpha(\delta) = 0.0328$. ◀

Randomized Tests

It has been emphasized by some statisticians that $\alpha(\delta)$ can be made exactly 0.05 in Example 9.2.6 if a *randomized* test procedure is used. Such a procedure is described as follows: When the rejection region of the test procedure contains all values of y greater than 4, we found in Example 9.2.6 that the size of the test is $\alpha(\delta) = 0.0328$. Also, when the point $y = 4$ is added to this rejection region, the value of $\alpha(\delta)$ jumps to 0.1209. Suppose, however, that instead of choosing between including the point $y = 4$ in the rejection region and excluding that point, we use an auxiliary randomization to decide whether or not to reject H_0 when $y = 4$. For example, we may toss a coin or spin a wheel to arrive at this decision. Then, by choosing appropriate probabilities to be used in this randomization, we can make $\alpha(\delta)$ exactly 0.05.

Specifically, consider the following test procedure: The hypothesis H_0 is rejected if $y > 4$, and H_0 is not rejected if $y < 4$. However, if $y = 4$, then an auxiliary randomization is carried out in which H_0 will be rejected with probability 0.195, and H_0 will not be rejected with probability 0.805. The size $\alpha(\delta)$ of this test will then be

$$\begin{aligned}\alpha(\delta) &= \Pr(Y > 4 | p = 0.2) + (0.195) \Pr(Y = 4 | p = 0.2) \\ &= 0.0328 + (0.195)(0.0881) = 0.05.\end{aligned} \tag{9.2.24}$$

Randomized tests do not seem to have any place in practical applications of statistics. It does not seem reasonable for a statistician to decide whether or not to reject a null hypothesis by tossing a coin or performing some other type of randomization for the sole purpose of obtaining a value of $\alpha(\delta)$ that is equal to some arbitrarily specified value such as 0.05. The main consideration for the statistician is to use a nonrandomized test procedure δ' having the form specified in the Nayman-Pearson lemma.

The proofs of Theorems 9.2.1 and 9.2.2 can be extended to find optimal tests among all tests regardless of whether they are randomized or nonrandomized. The optimal test in the extension of Theorem 9.2.2 has the same form as δ^* except that randomization is allowed whenever $f_1(\boldsymbol{x}) = kf_0(\boldsymbol{x})$. The only real need for randomized tests, in this book, will be the simplification that they provide for one step in the proof of Theorem 9.3.1 (page 562).

Furthermore, rather than fixing a specific size $\alpha(\delta)$ and trying to minimize $\beta(\delta)$, it might be more reasonable for the statistician to minimize a linear combination of the form $a\alpha(\delta) + b\beta(\delta)$. As we have seen in Theorem 9.2.1, such a minimization can always be achieved without recourse to an auxiliary randomization. In Sec. 9.9, we shall present another argument that indicates why it might be more reasonable to minimize a linear combination of the form $a\alpha(\delta) + b\beta(\delta)$ than to specify a value of $\alpha(\delta)$ and then minimize $\beta(\delta)$.

Summary

For the special case in which there are only two possible values, θ_0 and θ_1, for the parameter, we found a collection of procedures for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ that contains the optimal test procedure for each of the following criteria:

- Choose the test δ with the smallest value of $a\alpha(\delta) + b\beta(\delta)$.
- Among all tests δ with $\alpha(\delta) \le \alpha_0$, choose the test with the smallest value of $\beta(\delta)$.

Here, $\alpha(\delta) = \Pr(\text{Reject } H_0 | \theta = \theta_0)$ and $\beta(\delta) = \Pr(\text{Don't Reject } H_0 | \theta = \theta_1)$ are, respectively, the probabilities of type I and type II errors. The tests all have the following form for some positive constant k: reject H_0 if $f_0(\boldsymbol{x}) < kf_1(\boldsymbol{x})$, don't reject H_0 if $f_0(\boldsymbol{x}) > kf_1(\boldsymbol{x})$, and do either if $f_0(\boldsymbol{x}) = kf_1(\boldsymbol{x})$.

Exercises

1. Let $f_0(x)$ be the p.f. of the Bernoulli distribution with parameter 0.3, and let $f_1(x)$ be the p.f. of the Bernoulli distribution with parameter 0.6. Suppose that a single observation X is taken from a distribution for which the p.d.f. $f(x)$ is either $f_0(x)$ or $f_1(x)$, and the following simple hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad f(x) = f_0(x), \\ H_1&: \quad f(x) = f_1(x). \end{aligned}$$

Find the test procedure δ for which the value of $\alpha(\delta) + \beta(\delta)$ is a minimum.

2. Consider two p.d.f.'s $f_0(x)$ and $f_1(x)$ that are defined as follows:

$$f_0(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_1(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that a single observation X is taken from a distribution for which the p.d.f. $f(x)$ is either $f_0(x)$ or $f_1(x)$, and the following simple hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad f(x) = f_0(x), \\ H_1&: \quad f(x) = f_1(x). \end{aligned}$$

a. Describe a test procedure for which the value of $\alpha(\delta) + 2\beta(\delta)$ is a minimum.

b. Determine the minimum value of $\alpha(\delta) + 2\beta(\delta)$ attained by that procedure.

3. Consider again the conditions of Exercise 2, but suppose now that it is desired to find a test procedure for which the value of $3\alpha(\delta) + \beta(\delta)$ is a minimum.

a. Describe the procedure.

b. Determine the minimum value of $3\alpha(\delta) + \beta(\delta)$ attained by the procedure.

4. Consider again the conditions of Exercise 2, but suppose now that it is desired to find a test procedure for which $\alpha(\delta) \le 0.1$ and $\beta(\delta)$ is a minimum.

a. Describe the procedure.

b. Determine the minimum value of $\beta(\delta)$ attained by the procedure.

5. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean θ and known variance is 1, and the following hypotheses are to be tested:

$$H_0: \quad \theta = 3.5,$$
$$H_1: \quad \theta = 5.0.$$

a. Among all test procedures for which $\beta(\delta) \le 0.05$, describe a procedure for which $\alpha(\delta)$ is a minimum.

b. For $n = 4$, find the minimum value of $\alpha(\delta)$ attained by the procedure described in part (a).

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p. Let p_0 and p_1 be specified values such that $0 < p_1 < p_0 < 1$, and suppose that it is desired to test the following simple hypotheses:

$$H_0: \quad p = p_0,$$
$$H_1: \quad p = p_1.$$

a. Show that a test procedure for which $\alpha(\delta) + \beta(\delta)$ is a minimum rejects H_0 when $\overline{X}_n < c$.

b. Find the value of the constant c.

7. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean μ and unknown variance σ^2, and the following simple hypotheses are to be tested:

$$H_0: \quad \sigma^2 = 2,$$
$$H_1: \quad \sigma^2 = 3.$$

a. Show that among all test procedures for which $\alpha(\delta) \le 0.05$, the value of $\beta(\delta)$ is minimized by a procedure that rejects H_0 when $\sum_{i=1}^{n}(X_i - \mu)^2 > c$.

b. For $n = 8$, find the value of the constant c that appears in part (a).

8. Suppose that a single observation X is taken from the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown, and the following simple hypotheses are to be tested:

$$H_0: \quad \theta = 1,$$
$$H_1: \quad \theta = 2.$$

a. Show that there exists a test procedure for which $\alpha(\delta) = 0$ and $\beta(\delta) < 1$.

b. Among all test procedures for which $\alpha(\delta) = 0$, find the one for which $\beta(\delta)$ is a minimum.

9. Suppose that a random sample $X_1, \ldots, X_n$ is drawn from the uniform distribution on the interval $[0, \theta]$, and consider again the problem of testing the simple hypotheses described in Exercise 8. Find the minimum value of $\beta(\delta)$ that can be attained among all test procedures for which $\alpha(\delta) = 0$.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with unknown mean λ. Let λ_0 and λ_1 be specified values such that $\lambda_1 > \lambda_0 > 0$, and suppose that it is desired to test the following simple hypotheses:

$$H_0: \quad \lambda = \lambda_0,$$
$$H_1: \quad \lambda = \lambda_1.$$

a. Show that the value of $\alpha(\delta) + \beta(\delta)$ is minimized by a test procedure which rejects H_0 when $\overline{X}_n > c$.

b. Find the value of c.

c. For $\lambda_0 = 1/4$, $\lambda_1 = 1/2$, and $n = 20$, determine the minimum value of $\alpha(\delta) + \beta(\delta)$ that can be attained.

11. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known standard deviation 2, and the following simple hypotheses are to be tested:

$$H_0: \quad \mu = -1,$$
$$H_1: \quad \mu = 1.$$

Determine the minimum value of $\alpha(\delta) + \beta(\delta)$ that can be attained for each of the following values of the sample size n:

a. $n = 1$ **b.** $n = 4$ **c.** $n = 16$ **d.** $n = 36$

12. Let $X_1, \ldots, X_n$ be a random sample from the exponential distribution with unknown parameter θ. Let $0 < \theta_0 < \theta_1$ be two possible values of the parameter. Suppose that we wish to test the following hypotheses:

$$H_0: \quad \theta = \theta_0,$$
$$H_1: \quad \theta = \theta_1.$$

For each $\alpha_0 \in (0, 1)$, show that among all tests δ satisfying $\alpha(\delta) \le \alpha_0$, the test with the smallest probability of type II error will reject H_0 if $\sum_{i=1}^{n} X_i < c$, where c is the α_0 quantile of the gamma distribution with parameters n and θ_0.

13. Consider the series of examples in this section concerning service times in a queue. Suppose that the manager observes two service times X_1 and X_2. It is easy to see that both $f_1(\boldsymbol{x})$ in (9.2.1) and $f_0(\boldsymbol{x})$ in (9.2.2) depend on the observed data only through the value $t = x_1 + x_2$ of the statistic $T = X_1 + X_2$. Hence, the tests from Theorems 9.2.1 and 9.2.2 both depend only on the value of T.

a. Using Theorem 9.2.1, determine the test procedure that minimizes the sum of the probabilities of type I and type II errors.

b. Suppose that $X_1 = 4$ and $X_2 = 3$ are observed. Perform the test in part (a) to see whether H_0 is rejected.

c. Prove that the distribution of T, given that H_0 is true, is the gamma distribution with parameters 2 and 1/2.

d. Using Theorem 9.2.2, determine the test procedure with level at most 0.01 that has minimum probability of type II error. *Hint:* It looks like you need to solve a system of nonlinear equations, but for a level 0.01 test, the equations collapse to a single simple equation.

e. Suppose that $X_1 = 4$ and $X_2 = 3$ are observed. Perform the test in part (d) to see whether H_0 is rejected.

⋆ 9.3 Uniformly Most Powerful Tests

When the null and/or alternative hypothesis is composite, we can still find a class of tests that has optimal properties in certain circumstances. In particular, the null and alternative hypotheses must be of the form $H_0 : \theta \le \theta_0$ and $H_1 : \theta > \theta_0$, or $H_0 : \theta \ge \theta_0$ and $H_1 : \theta < \theta_0$. In addition, the family of distributions of the data must have a property called "monotone likelihood ratio," which is defined in this section.

Definition of a Uniformly Most Powerful Test

Example 9.3.1

Service Times in a Queue. In Example 9.2.1, a manager was interested in testing which of two joint distributions described the service times in a queue that she was managing. Suppose, now, that instead of considering only two joint distributions, the manager wishes to consider all of the joint distributions that can be described by saying that the service times form a random sample from the exponential distribution with parameter θ conditional on θ. That is, for each possible rate $\theta > 0$, the manager is willing to consider the possibility that the service times are i.i.d. exponential random variables with parameter θ. In particular, the manager is interested in testing $H_0 : \theta \le 1/2$ versus $H_1 : \theta > 1/2$. For each $\theta' > 1/2$, the manager could use the methods of Sec. 9.2 to test the hypotheses $H_0' : \theta = 1/2$ versus $H_1' : \theta = \theta'$. She could obtain the level α_0 test with the smallest possible type II error probability when $\theta = \theta'$. But can she find a single level α_0 test that has the largest possible type II error probability simultaneously for all $\theta > 1/2$? And will that test have probability of type I error at most α_0 for all $\theta \le 1/2$? ◀

Consider a problem of testing hypotheses in which the random variables $\boldsymbol{X} = (X_1, \dots, X_n)$ form a random sample from a distribution for which either the p.d.f. or the p.f. is $f(x|\theta)$. We suppose that the value of the parameter θ is unknown but must lie in a specified parameter space Ω that is a subset of the real line. As usual, we shall suppose that Ω_0 and Ω_1 are disjoint subsets of Ω, and the hypotheses to be tested are

$$\begin{aligned} H_0&: \quad \theta \in \Omega_0, \\ H_1&: \quad \theta \in \Omega_1. \end{aligned} \tag{9.3.1}$$

We shall assume that the subset Ω_1 contains at least two distinct values of θ, in which case the alternative hypothesis H_1 is composite. The null hypothesis H_0 may be either simple or composite. Example 9.3.1 is of the type just described with $\Omega_0 = (0, 1/2]$ and $\Omega_1 = (1/2, \infty)$.

We shall also suppose that it is desired to test the hypotheses (9.3.1) at a specified level of significance α_0, where α_0 is a given number in the interval $0 < \alpha_0 < 1$. In other words, we shall consider only procedures in which $\Pr(\text{Rejecting } H_0|\theta) \leq \alpha_0$ for every value of $\theta \in \Omega_0$. If $\pi(\theta|\delta)$ denotes the power function of a given test procedure δ, this requirement can be written simply as

$$\pi(\theta|\delta) \leq \alpha_0 \quad \text{for } \theta \in \Omega_0. \tag{9.3.2}$$

Equivalently, if $\alpha(\delta)$ denotes the size of a test procedure δ, as defined by Eq. (9.1.7), then the requirement (9.3.2) can also be expressed by the relation

$$\alpha(\delta) \leq \alpha_0. \tag{9.3.3}$$

Finally, among all test procedures that satisfy the requirement (9.3.3), we want to find one that has the smallest possible probability of type II error for every $\theta \in \Omega_1$. In terms of the power function, we want the value of $\pi(\theta|\delta)$ to be as large as possible for every value of $\theta \in \Omega_1$.

It may not be possible to satisfy this last criterion. If θ_1 and θ_2 are two different values of θ in Ω_1, then the test procedure for which the value of $\pi(\theta_1|\delta)$ is a maximum might be different from the test procedure for which the value of $\pi(\theta_2|\delta)$ is a maximum. In other words, there might be no single test procedure δ that maximizes the power function $\pi(\theta|\delta)$ simultaneously for every value of θ in Ω_1. In some problems, however, there will exist a test procedure that satisfies this criterion. Such a procedure, when it exists, is called a *uniformly most powerful* test, or, more briefly, a UMP test. The formal definition of a UMP test is as follows.

Definition 9.3.1 Uniformly Most Powerful (UMP) Test. A test procedure δ^* is a *uniformly most powerful (UMP) test* of the hypotheses (9.3.1) at the level of significance α_0 if $\alpha(\delta^*) \leq \alpha_0$ and, for every other test procedure δ such that $\alpha(\delta) \leq \alpha_0$, it is true that

$$\pi(\theta|\delta) \leq \pi(\theta|\delta^*) \quad \text{for every value of } \theta \in \Omega_1. \tag{9.3.4}$$

In this section, we shall show that a UMP test exists in many problems in which the random sample comes from one of the standard families of distributions that we have been considering in this book.

Monotone Likelihood Ratio

Example 9.3.2 Service Times in a Queue. Suppose that the manager in Example 9.3.1 observes a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of service times and tries to find the level α_0 test of $H_0' : \theta = 1/2$ versus $H_1' : \theta = \theta'$ that has the largest power at $\theta = \theta' > 1/2$. According to Exercise 12 in Sec. 9.2, the test will reject H_0' if $\sum_{i=1}^{n} X_i$ is less than the α_0 quantile of the gamma distribution with parameters n and 1/2. This test is the same test regardless of which $\theta' > 1/2$ the manager considers. Hence, the test is UMP at the level of significance α_0 for testing $H_0' : \theta = 1/2$ versus $H_1 : \theta > 1/2$. ◀

The family of exponential distributions in Example 9.3.2 has a special property called *monotone likelihood ratio* that allows the manager to find a UMP test.

Definition 9.3.2 Monotone Likelihood Ratio. Let $f_n(\boldsymbol{x}|\theta)$ denote the joint p.d.f. or the joint p.f. of the observations $\boldsymbol{X} = (X_1, \ldots, X_n)$. Let $T = r(\boldsymbol{X})$ be a statistic. It is said that the joint distribution of $\boldsymbol{X}$ has a *monotone likelihood ratio (MLR) in the statistic T* if the following property is satisfied: For every two values $\theta_1 \in \Omega$ and $\theta_2 \in \Omega$, with $\theta_1 < \theta_2$, the ratio $f_n(\boldsymbol{x}|\theta_2)/f_n(\boldsymbol{x}|\theta_1)$ depends on the vector $\boldsymbol{x}$ only through the function $r(\boldsymbol{x})$, and this ratio is a monotone function of $r(\boldsymbol{x})$ over the range of possible values of $r(\boldsymbol{x})$. Specifically, if the ratio is increasing, we say that the distribution of $\boldsymbol{X}$ has *increasing MLR*, and if the ratio is decreasing, we say that the distribution has *decreasing MLR*.

Example 9.3.3 **Sampling from a Bernoulli Distribution.** Suppose that $X_1, \dots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p $(0 < p < 1)$. If we let $y = \sum_{i=1}^n x_i$, then the joint p.f. $f_n(\boldsymbol{x}|p)$ is as follows:

$$f_n(\boldsymbol{x}|p) = p^y(1-p)^{n-y}.$$

Therefore, for every two values p_1 and p_2 such that $0 < p_1 < p_2 < 1$,

$$\frac{f_n(\boldsymbol{x}|p_2)}{f_n(\boldsymbol{x}|p_1)} = \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^y \left(\frac{1-p_2}{1-p_1}\right)^n. \tag{9.3.5}$$

It can be seen from Eq. (9.3.5) that the ratio $f_n(\boldsymbol{x}|p_2)/f_n(\boldsymbol{x}|p_1)$ depends on the vector $\boldsymbol{x}$ only through the value of y, and this ratio is an increasing function of y. Therefore, $f_n(\boldsymbol{x}|p)$ has increasing monotone likelihood ratio in the statistic $Y = \sum_{i=1}^n X_i$. ◄

Example 9.3.4 **Sampling from an Exponential Distribution.** Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random sample from the exponential distribution with unknown parameter $\theta > 0$. The joint p.d.f. is

$$f_n(\boldsymbol{x}|\theta) = \begin{cases} \theta^n \exp\left(-\theta \sum_{i=1}^n x_i\right) & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

For $0 < \theta_1 < \theta_2$, we have

$$\frac{f_n(\boldsymbol{x}|\theta_2)}{f_n(\boldsymbol{x}|\theta_1)} = \left(\frac{\theta_2}{\theta_1}\right)^n \exp\left([\theta_1 - \theta_2]\sum_{i=1}^n x_i\right), \tag{9.3.6}$$

if all $x_i > 0$. If we let $r(\boldsymbol{x}) = \sum_{i=1}^n x_i$, then we see that the ratio in Eq. (9.3.6) depends on $\boldsymbol{x}$ only through $r(\boldsymbol{x})$ and is a decreasing function of $r(\boldsymbol{x})$. Hence, the joint distribution of a random sample of exponential random variables has decreasing MLR in $T = \sum_{i=1}^n X_i$. ◄

In Example 9.3.4, we could have defined the statistic $T' = -\sum_{i=1}^n X_i$ or $T' = 1/\sum_{i=1}^n X_i$, and then the distribution would have had increasing MLR in T'. This can be done in general in Definition 9.3.2. For this reason, when we prove theorems that assume that a distribution has MLR, we shall state and prove the theorems for increasing MLR only. When a distribution has decreasing MLR, the reader can transform the statistic by a strictly decreasing function and then transform the result back to the original statistic, if desired.

Example 9.3.5 **Sampling from a Normal Distribution.** Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ $(-\infty < \mu < \infty)$ and known variance σ^2. The joint p.d.f. $f_n(\boldsymbol{x}|\mu)$ is as follows:

$$f_n(\boldsymbol{x}|\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right].$$

Therefore, for every two values μ_1 and μ_2 such that $\mu_1 < \mu_2$,

$$\frac{f_n(\boldsymbol{x}|\mu_2)}{f_n(\boldsymbol{x}|\mu_1)} = \exp\left\{\frac{n(\mu_2 - \mu_1)}{\sigma^2}\left[\overline{x}_n - \frac{1}{2}(\mu_2 + \mu_1)\right]\right\}. \tag{9.3.7}$$

It can be seen from Eq. (9.3.7) that the ratio $f_n(\boldsymbol{x}|\mu_2)/f_n(\boldsymbol{x}|\mu_1)$ depends on the vector $\boldsymbol{x}$ only through the value of $\overline{x}_n$, and this ratio is an increasing function of $\overline{x}_n$. Therefore, $f_n(\boldsymbol{x}|\mu)$ has increasing monotone likelihood ratio in the statistic $\overline{X}_n$. ◄

One-Sided Alternatives

In Example 9.3.2, we found a UMP level α_0 test for a simple null hypothesis $H_0' : \theta = 1/2$ against a one-sided alternative $H_1 : \theta > 1/2$. It is more common in such problems to test hypotheses of the form

$$\begin{aligned} H_0&: \quad \theta \leq \theta_0, \\ H_1&: \quad \theta > \theta_0. \end{aligned} \tag{9.3.8}$$

That is, both the null and alternative hypotheses are one-sided. Because the one-sided null hypothesis is larger than the simple null $H_0' : \theta = \theta_0$, it is not necessarily the case that a level α_0 test of H_0' will be a level α_0 test of H_0. However, if the joint distribution of the observations has MLR, we will be able to show that there will exist UMP level α_0 tests of the hypotheses (9.3.8). Furthermore (see Exercise 12), there will exist UMP tests of the hypotheses obtained by reversing the inequalities in both H_0 and H_1 in (9.3.8).

Theorem 9.3.1 Suppose that the joint distribution of $\boldsymbol{X}$ has increasing monotone likelihood ratio in the statistic $T = r(\boldsymbol{X})$. Let c and α_0 be constants such that

$$\Pr(T \geq c|\theta = \theta_0) = \alpha_0. \tag{9.3.9}$$

Then the test procedure δ^* that rejects H_0 if $T \geq c$ is a UMP test of the hypotheses (9.3.8) at the level of significance α_0. Also, $\pi(\theta|\delta^*)$ is a monotone increasing function of θ.

Proof Let $\theta' < \theta''$ be arbitrary values of θ. Let $\alpha_0' = \pi(\theta'|\delta^*)$. It follows from the Nayman-Pearson lemma that among all procedures δ for which

$$\pi(\theta'|\delta) \leq \alpha_0', \tag{9.3.10}$$

the value of $\pi(\theta''|\delta)$ will be maximized $(1 - \pi(\theta''|\delta)$ minimized) by a procedure that rejects H_0 when $f_n(\boldsymbol{x}|\theta'')/f_n(\boldsymbol{x}|\theta') \geq k$. The constant k is to be chosen so that

$$\pi(\theta'|\delta) = \alpha_0'. \tag{9.3.11}$$

Because the distribution of $\boldsymbol{X}$ has increasing MLR, the likelihood ratio $f_n(\boldsymbol{x}|\theta'')/f_n(\boldsymbol{x}|\theta')$ is an increasing function of $r(\boldsymbol{x})$. Therefore, a procedure that rejects H_0 when the likelihood ratio is at least equal to k will be equivalent to a procedure that rejects H_0 when $r(\boldsymbol{x})$ is at least equal to some other number c. The value of c is to be chosen so that (9.3.11) holds. The test δ^* satisfies Eq. (9.3.11) and has the correct form; hence, it maximizes the power function at $\theta = \theta''$ among all tests that satisfy Eq. (9.3.10). Another test δ that satisfies Eq. (9.3.10) is the following: Flip a coin that has probability of heads equal to α_0', and reject H_0 if the coin lands heads. This test has $\pi(\theta|\delta) = \alpha_0'$ for all θ including θ' and θ''. Because δ^* maximizes the power function at θ'', we have

$$\pi(\theta''|\delta^*) \geq \pi(\theta'|\delta) = \alpha_0' = \pi(\theta'|\delta^*). \tag{9.3.12}$$

Hence, we have proven the claim that $\pi(\theta|\delta^*)$ is a monotone increasing function of θ.

Next, consider the special case of what we have just proven with $\theta' = \theta_0$. Then $\alpha_0' = \alpha_0$, and we have proven that, for every $\theta'' > \theta_0$, δ^* maximizes $\pi(\theta''|\delta)$ among all tests δ that satisfy

$$\pi(\theta_0|\delta) \le \alpha_0. \tag{9.3.13}$$

Every level α_0 test δ satisfies Eq. (9.3.13). Hence, δ^* has power at θ'' at least as high as the power of every level α_0 test. All that remains to complete the proof is to show that δ^* is itself a level α_0 test.

We have already shown that the power function $\pi(\theta|\delta^*)$ is monotone increasing. Hence, $\pi(\theta|\delta^*) \le \alpha_0$ for all $\theta \le \theta_0$, and δ^* is a level α_0 test. ■

Example 9.3.6 Service Times in a Queue. The manager in Example 9.3.2 might be interested in the hypotheses $H_0: \theta \le 1/2$ versus $H_1: \theta > 1/2$. The distribution in that example has decreasing MLR in the statistic $T = \sum_{i=1}^{n} X_i$, and hence it has increasing MLR in $-T$. Theorem 9.3.1 says that a UMP level α_0 test is to reject H_0 when $-T$ is greater than the $1-\alpha_0$ quantile of the distribution of $-T$ given $\theta = 1/2$. This is the same as rejecting H_0 when T is less than the α_0 quantile of the distribution of T. The distribution of T given $\theta = 1/2$ is the gamma distribution with parameters n and 1/2, which is also the χ^2 distribution with $2n$ degrees of freedom. For example, if $n = 10$ and $\alpha_0 = 0.1$, the quantile is 12.44, which can be found in the table in the back of the book or from computer software. ◄

Example 9.3.7 Testing Hypotheses about the Proportion of Defective Items. Suppose that the proportion p of defective items in a large manufactured lot is unknown, 20 items are to be selected at random from the lot and inspected, and the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad p \le 0.1, \\ H_1&: \quad p > 0.1. \end{aligned} \tag{9.3.14}$$

We shall show first that there exist UMP tests of the hypotheses (9.3.14). We shall then determine the form of these tests and discuss the different levels of significance that can be attained with nonrandomized tests.

Let $X_1, \dots, X_{20}$ denote the 20 random variables in the sample. Then $X_1, \dots, X_{20}$ form a random sample of size 20 from the Bernoulli distribution with parameter p, and it is known from Example 9.3.3 that the joint p.f. of $X_1, \dots, X_{20}$ has increasing monotone likelihood ratio in the statistic $Y = \sum_{i=1}^{20} X_i$. Therefore, by Theorem 9.3.1, a test procedure that rejects H_0 when $Y \ge c$ will be a UMP test of the hypotheses (9.3.14).

For each specific choice of the constant c, the size of the UMP test will be $\alpha_0 = \Pr(Y \ge c|p = 0.1)$. When $p = 0.1$, the random variable Y has the binomial distribution with parameters $n = 20$ and $p = 0.1$. Because Y has a discrete distribution and assumes only a finite number of different possible values, it follows that there are only a finite number of different possible values for α_0. To illustrate this remark, it is found from a table of the binomial distribution that if $c = 7$, then $\alpha_0 = \Pr(Y \ge 7|p = 0.1) = 0.0024$, and if $c = 6$, then $\alpha_0 = \Pr(Y \ge 6|p = 0.1) = 0.0113$. Therefore, if an experimenter wants the size of the test to be approximately 0.01, she could choose either $c = 7$ and $\alpha_0 = 0.0024$ or $c = 6$ and $\alpha_0 = 0.0113$. The test with $c = 7$ is a level 0.01 test while the test with $c = 6$ is not, because the size of the former test is less than 0.01 while the size of the latter test is greater than 0.01.

If the experimenter wants the size of the test to be exactly 0.01, then she can use a randomized test procedure of the type described in Sec. 9.2. ◄

Example 9.3.8 Testing Hypotheses about the Mean of a Normal Distribution. Let $X_1, \dots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Assume that σ^2 is known. Let μ_0 be a specified number, and suppose that the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \mu \le \mu_0, \\ H_1&: \quad \mu > \mu_0. \end{aligned} \tag{9.3.15}$$

We shall show first that, for every specified level of significance α_0 $(0 < \alpha_0 < 1)$, there is a UMP test of the hypotheses (9.3.15) with size equal to α_0. We shall then determine the power function of the UMP test.

It is known from Example 9.3.5 that the joint p.d.f. of $X_1, \ldots, X_n$ has an increasing monotone likelihood ratio in the statistic $\overline{X}_n$. Therefore, by Theorem 9.3.1, a test procedure δ_1 that rejects H_0 when $\overline{X}_n \ge c$ is a UMP test of the hypotheses (9.3.15). The size of this test is $\alpha_0 = \Pr(\overline{X}_n \ge c|\mu = \mu_0)$.

Since $\overline{X}_n$ has a continuous distribution, c is the $1 - \alpha_0$ quantile of the distribution of $\overline{X}_n$ given $\mu = \mu_0$. That is, c is the $1 - \alpha_0$ quantile of the normal distribution with mean μ_0 and variance σ^2/n. As we learned in Chapter 5, this quantile is

$$c = \mu_0 + \Phi^{-1}(1 - \alpha_0)\sigma n^{-1/2}, \tag{9.3.16}$$

where Φ^{-1} is the quantile function of the standard normal distribution. For simplicity, we shall let $z_{\alpha_0} = \Phi^{-1}(1 - \alpha_0)$ for the rest of this example.

We shall now determine the power function $\pi(\mu|\delta_1)$ of this UMP test. By definition,

$$\pi(\mu|\delta_1) = \Pr(\text{Rejecting } H_0|\mu) = \Pr(\overline{X}_n \ge \mu_0 + z_{\alpha_0}\sigma n^{-1/2}|\mu). \tag{9.3.17}$$

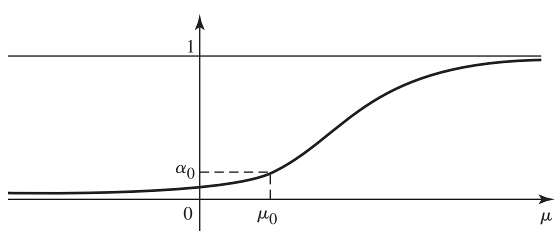
For every value of μ, the random variable $Z' = n^{1/2}(\overline{X}_n - \mu)/\sigma$ will have the standard normal distribution. Therefore, if Φ denotes the c.d.f. of the standard normal distribution, then

$$\begin{aligned} \pi(\mu|\delta_1) &= \Pr\left[Z' \ge z_{\alpha_0} + \frac{n^{1/2}(\mu_0 - \mu)}{\sigma} \right] \\ &= 1 - \Phi\left[z_{\alpha_0} + \frac{n^{1/2}(\mu_0 - \mu)}{\sigma} \right] = \Phi\left[\frac{n^{1/2}(\mu - \mu_0)}{\sigma} - z_{\alpha_0} \right]. \end{aligned} \tag{9.3.18}$$

The power function $\pi(\mu|\delta_1)$ is sketched in Fig. 9.6. ◄

In each of the pairs of hypotheses (9.3.8), (9.3.14), and (9.3.15), the alternative hypothesis H_1 is called a *one-sided alternative* because the set of possible values of the parameter under H_1 lies entirely on one side of the set of possible values under the null hypothesis H_0. In particular, for the hypotheses (9.3.8), (9.3.14), or (9.3.15), every possible value of the parameter under H_1 is larger than every possible value under H_0.

Figure 9.6 The power function $\pi(\mu|\delta_1)$ for the UMP test of the hypotheses (9.3.15).



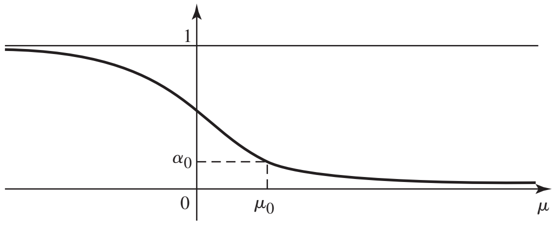


Figure 9.7 The power function $\pi(\mu|\delta_2)$ for the UMP test of the hypotheses (9.3.19).

Example 9.3.9

One-Sided Alternatives in the Other Direction. Suppose now that instead of testing the hypotheses (9.3.15) in Example 9.3.8, we are interested in testing the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu \geq \mu_0, \\ H_1&: \quad \mu < \mu_0. \end{aligned} \tag{9.3.19}$$

In this case, the hypothesis H_1 is again a one-sided alternative, and it can be shown (see Exercise 12) that there exists a UMP test of the hypotheses (9.3.19) at every specified level of significance α_0 $(0 < \alpha_0 < 1)$. By analogy with Eq. (9.3.16), the UMP test δ_2 will reject H_0 when $\overline{X}_n \leq c$, where

$$c = \mu_0 - \Phi^{-1}(1 - \alpha_0)\sigma n^{-1/2}. \tag{9.3.20}$$

The power function $\pi(\mu|\delta_2)$ of the test δ_2 will be

$$\pi(\mu|\delta_2) = \Pr(\overline{X}_n \leq c|\mu) = \Phi\left[\frac{n^{1/2}(\mu_0 - \mu)}{\sigma} - \Phi^{-1}(1 - \alpha_0)\right]. \tag{9.3.21}$$

This function is sketched in Fig. 9.7. Indeed, Exercise 12 extends Theorem 9.3.1 to one-sided hypotheses of the form (9.3.19) in every monotone likelihood ratio family. In Sec. 9.8, we shall show that for all one-sided cases with monotone likelihood ratio, the tests of the form given in Theorem 9.3.1 and Exercise 12 are also optimal when one focuses on the posterior distribution of θ rather than on the power function. ◀

Two-Sided Alternatives

Suppose, finally, that instead of testing either the hypotheses (9.3.15) in Example 9.3.8 or the hypotheses (9.3.19), we are interested in testing the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu = \mu_0, \\ H_1&: \quad \mu \neq \mu_0. \end{aligned} \tag{9.3.22}$$

In this case, H_0 is a simple hypothesis and H_1 is a two-sided alternative. Since H_0 is a simple hypothesis, the size of every test procedure δ will simply be equal to the value $\pi(\mu_0|\delta)$ of the power function at the point $\mu = \mu_0$.

Indeed, for each α_0 $(0 < \alpha_0 < 1)$, there is no UMP test of the hypotheses (9.3.22) at level of significance α_0. For every value of μ such that $\mu > \mu_0$, the value of $\pi(\mu|\delta)$ will be maximized by the test procedure δ_1 in Example 9.3.8, whereas for every value of μ such that $\mu < \mu_0$, the value of $\pi(\mu|\delta)$ will be maximized by the test procedure δ_2 in Example 9.3.9. It can be shown (see Exercise 19) that δ_1 is essentially the unique test that maximizes $\pi(\mu|\delta)$ for $\mu > \mu_0$. Since δ_1 does not maximize $\pi(\mu|\delta)$ for $\mu < \mu_0$, no test could maximize $\pi(\mu|\delta)$ simultaneously for $\mu > \mu_0$ and $\mu < \mu_0$. In the next section, we shall discuss the selection of an appropriate test procedure in this problem.

Summary

A uniformly most powerful (UMP) level α_0 test is a level α_0 test whose power function on the alternative hypothesis is always at least as high as the power function of every level α_0 test. If the family of distributions for the data has a monotone likelihood ratio in a statistic T, and if the null and alternative hypotheses are both one-sided, then there exists a UMP level α_0 test. In these cases, the UMP level α_0 test is either of the form "reject H_0 if $T \ge c$" or "reject H_0 if $T \le c$."

Exercises

1. Suppose that $X_1, \dots, X_n$ form a random sample from the Poisson distribution with unknown mean λ ($\lambda > 0$). Show that the joint p.f. of $X_1, \dots, X_n$ has a monotone likelihood ratio in the statistic $\sum_{i=1}^n X_i$.

2. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with known mean μ and unknown variance σ^2 ($\sigma^2 > 0$). Show that the joint p.d.f. of $X_1, \dots, X_n$ has a monotone likelihood ratio in the statistic $\sum_{i=1}^n (X_i - \mu)^2$.

3. Suppose that $X_1, \dots, X_n$ form a random sample from the gamma distribution with parameters α and β. Assume that α is unknown ($\alpha > 0$) and that β is known. Show that the joint p.d.f. of $X_1, \dots, X_n$ has a monotone likelihood ratio in the statistic $\prod_{i=1}^n X_i$.

4. Suppose that $X_1, \dots, X_n$ form a random sample from the gamma distribution with parameters α and β. Assume that α is known and that β is unknown ($\beta > 0$). Show that the joint p.d.f. of $X_1, \dots, X_n$ has a monotone likelihood ratio in the statistic $-\overline{X}_n$.

5. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution that belongs to an exponential family, as defined in Exercise 23 of Sec. 7.3, and the p.d.f. or the p.f. of this distribution is $f(\boldsymbol{x}|\theta)$, as given in that exercise. Suppose also that $c(\theta)$ is a strictly increasing function of θ. Show that the joint p.d.f. or the joint p.f. of $X_1, \dots, X_n$ has a monotone likelihood ratio in the statistic $\sum_{i=1}^n d(X_i)$.

6. Suppose that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$. Show that the joint p.d.f. of $X_1, \dots, X_n$ has a monotone likelihood ratio in the statistic $\max\{X_1, \dots, X_n\}$.

7. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution involving a parameter θ whose value is unknown, and suppose that it is desired to test the following hypotheses:

$$H_0: \quad \theta \le \theta_0,$$
$$H_1: \quad \theta > \theta_0.$$

Suppose also that the test procedure to be used ignores the observed values in the sample and, instead, depends only on an auxiliary randomization in which an unbalanced coin is tossed so that a head will be obtained with probability 0.05, and a tail will be obtained with probability 0.95. If a head is obtained, then H_0 is rejected, and if a tail is obtained, then H_0 is not rejected. Describe the power function of this randomized test procedure.

8. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with known mean 0 and unknown variance σ^2, and suppose that it is desired to test the following hypotheses:

$$H_0: \quad \sigma^2 \le 2,$$
$$H_1: \quad \sigma^2 > 2.$$

Show that there exists a UMP test of these hypotheses at every level of significance α_0 ($0 < \alpha_0 < 1$).

9. Show that the UMP test in Exercise 8 rejects H_0 when $\sum_{i=1}^n X_i^2 \ge c$, and determine the value of c when $n = 10$ and $\alpha_0 = 0.05$.

10. Suppose that $X_1, \dots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p, and suppose that it is desired to test the following hypotheses:

$$H_0: \quad p \le \tfrac{1}{2},$$
$$H_1: \quad p > \tfrac{1}{2}.$$

Show that if the sample size is $n = 20$, then there exists a nonrandomized UMP test of these hypotheses at the level of significance $\alpha_0 = 0.0577$ and at the level of significance $\alpha_0 = 0.0207$.

11. Suppose that $X_1, \dots, X_n$ form a random sample from the Poisson distribution with unknown mean λ, and suppose that it is desired to test the following hypotheses:

$$H_0: \quad \lambda \le 1,$$
$$H_1: \quad \lambda > 1.$$

Show that if the sample size is $n = 10$, then there exists a nonrandomized UMP test of these hypotheses at the level of significance $\alpha_0 = 0.0143$.

12. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution that involves a parameter θ whose value is unknown, and the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x}|\theta)$ has a monotone likelihood ratio in the statistic $T = r(\boldsymbol{X})$. Let θ_0

be a specified value of θ, and suppose that the following hypotheses are to be tested:

$$H_0: \quad \theta \geq \theta_0,$$
$$H_1: \quad \theta < \theta_0.$$

Let c be a constant such that $\Pr(T \leq c|\theta = \theta_0) = \alpha_0$. Show that the test procedure which rejects H_0 if $T \leq c$ is a UMP test at the level of significance α_0.

13. Suppose that four observations are taken at random from the normal distribution with unknown mean μ and known variance 1. Suppose also that the following hypotheses are to be tested:

$$H_0: \quad \mu \geq 10,$$
$$H_1: \quad \mu < 10.$$

a. Determine a UMP test at the level of significance $\alpha_0 = 0.1$.

b. Determine the power of this test when $\mu = 9$.

c. Determine the probability of not rejecting H_0 if $\mu = 11$.

14. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with unknown mean λ, and suppose that it is desired to test the following hypotheses:

$$H_0: \quad \lambda \geq 1,$$
$$H_1: \quad \lambda < 1.$$

Suppose also that the sample size is $n = 10$. At what levels of significance α_0 in the interval $0 < \alpha_0 < 0.03$ do there exist nonrandomized UMP tests?

15. Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with unknown parameter β, and suppose that it is desired to test the following hypotheses:

$$H_0: \quad \beta \geq \tfrac{1}{2},$$
$$H_1: \quad \beta < \tfrac{1}{2}.$$

Show that at every level of significance α_0 $(0 < \alpha_0 < 1)$, there exists a UMP test that specifies rejecting H_0 when $\overline{X}_n \geq c$, for some constant c.

16. Consider again the conditions of Exercise 15, and suppose that the sample size is $n = 10$. Determine the value of the constant c that defines the UMP test at the level of significance $\alpha_0 = 0.05$. *Hint*: Use the table of the χ^2 distribution.

17. Consider a single observation X from the Cauchy distribution with unknown location parameter θ. That is, the p.d.f. of X is

$$f(x|\theta) = \frac{1}{\pi[1 + (x - \theta)^2]} \quad \text{for } -\infty < x < \infty.$$

Suppose that it is desired to test the following hypotheses:

$$H_0: \quad \theta = 0,$$
$$H_1: \quad \theta > 0.$$

Show that, for every α_0 $(0 < \alpha_0 < 1)$, there does not exist a UMP test of these hypotheses at level of significance α_0.

18. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance 1. Suppose also that the following hypotheses are to be tested:

$$H_0: \quad \mu \leq 0,$$
$$H_1: \quad \mu > 0.$$

Let δ^* denote the UMP test of these hypotheses at the level of significance $\alpha_0 = 0.025$, and let $\pi(\mu|\delta^*)$ denote the power function of δ^*.

a. Determine the smallest value of the sample size n for which $\pi(\mu|\delta^*) \geq 0.9$ for $\mu \geq 0.5$.

b. Determine the smallest value of n for which $\pi(\mu|\delta^*) \leq 0.001$ for $\mu \leq -0.1$.

19. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance σ^2. In this problem, you will prove the missing steps from the proof that there is no UMP level α_0 test for the hypotheses in (9.3.22). Let δ_1 be the test procedure with level α_0 defined in Example 9.3.8.

a. Let A be a set of possible values for the random vector $\boldsymbol{X} = (X_1 \ldots, X_n)$. Let $\mu_1 \neq \mu_0$. Prove that $\Pr(\boldsymbol{X} \in A|\mu = \mu_0) > 0$ if and only if $\Pr(\boldsymbol{X} \in A|\mu = \mu_1) > 0$.

b. Let δ be a size α_0 test for the hypotheses in (9.3.22) that differs from δ_1 in the following sense: There is a set A for which δ rejects its null hypothesis when $\boldsymbol{X} \in A$, δ_1 does not reject its null hypothesis when $\boldsymbol{X} \in A$, and $\Pr(\boldsymbol{X} \in A|\mu = \mu_0) > 0$. Prove that $\pi(\mu|\delta) < \pi(\mu|\delta_1)$ for all $\mu > \mu_0$.

★ 9.4 Two-Sided Alternatives

When testing a simple null hypothesis against a two-sided alternative (as at the end of Sec. 9.3), the choice of a test procedure requires a bit more care than in the one-sided case. This section discusses some of the issues and describes the most common choices.

General Form of the Procedure

Example 9.4.1

Egyptian Skulls. In Example 9.1.2, we considered how to compare measurements of skulls found in Egypt to modern measurements. For example, the average breadth of a modern-day skull is about 140mm. Suppose that we model the breadths of skulls from 4000 B.C. as normal random variables with unknown mean μ and known variance of 26. Unlike Example 9.1.6, suppose now that the researchers have no theory suggesting that skull breadths should increase over time. Instead, they are merely interested in whether breadths changed at all. How would they choose a test of the hypotheses $H_0 : \mu = 140$ versus $H_1 : \mu \neq 140$? ◀

In this section, we shall suppose that $\boldsymbol{X} = (X_1, \dots, X_n)$ is a random sample from a normal distribution for which the mean μ is unknown and the variance σ^2 is known, and that it is desired to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu = \mu_0, \\ H_1&: \quad \mu \neq \mu_0. \end{aligned} \tag{9.4.1}$$

In most practical problems, we would assume that both μ and σ^2 were unknown. We shall address that case in Sec. 9.5.

It was claimed at the end of Sec. 9.3 that there is no UMP test of the hypotheses (9.4.1) at any specified level of significance α_0 $(0 < \alpha_0 < 1)$. Neither the test procedure δ_1 nor the procedure δ_2 defined in Examples 9.3.8 and 9.3.9 is appropriate for testing the hypotheses (9.4.1), because each of those procedures has high power function only on one side of two-sided alternative H_1 and they each have low power function on the other side. However, the properties of the procedures δ_1 and δ_2 given in Sec. 9.3 and the fact that the sample mean $\overline{X}_n$ is the M.L.E. of μ suggest that a reasonable test of the hypotheses (9.4.1) would be to reject H_0 if $\overline{X}_n$ is far from μ_0. In other words, it seems reasonable to use a test procedure δ that rejects H_0 if either $\overline{X}_n \leq c_1$ or $\overline{X}_n \geq c_2$, where c_1 and c_2 are two suitably chosen constants, presumably with $c_1 < \mu_0$ and $c_2 > \mu_0$.

If the size of the test is to be α_0, then the values of c_1 and c_2 must be chosen so as to satisfy the following relation:

$$\Pr(\overline{X}_n \leq c_1 | \mu = \mu_0) + \Pr(\overline{X}_n \geq c_2 | \mu = \mu_0) = \alpha_0. \tag{9.4.2}$$

There are an infinite number of pairs of values of c_1 and c_2 that satisfy Eq. (9.4.2). When $\mu = \mu_0$, the random variable $n^{1/2}(\overline{X}_n - \mu_0)/\sigma$ has the standard normal distribution. If, as usual, we let Φ denote the c.d.f. of the standard normal distribution, then it follows that Eq. (9.4.2) is equivalent to the following relation:

$$\Phi\left[\frac{n^{1/2}(c_1 - \mu_0)}{\sigma}\right] + 1 - \Phi\left[\frac{n^{1/2}(c_2 - \mu_0)}{\sigma}\right] = \alpha_0. \tag{9.4.3}$$

Corresponding to every pair of positive numbers α_1 and α_2 such that $\alpha_1 + \alpha_2 = \alpha_0$, there exists a pair of numbers c_1 and c_2 such that $\Phi[n^{1/2}(c_1 - \mu_0)/\sigma] = \alpha_1$ and $1 - \Phi[n^{1/2}(c_2 - \mu_0)/\sigma] = \alpha_2$. Every such pair of values of c_1 and c_2 will satisfy Eqs. (9.4.2) and (9.4.3).

For example, suppose that $\alpha_0 = 0.05$. Then, choosing $\alpha_1 = 0.025$ and $\alpha_2 = 0.025$ yields a test procedure δ_3, which is defined by the values $c_1 = \mu_0 - 1.96\sigma n^{-1/2}$ and $c_2 = \mu_0 + 1.96\sigma n^{-1/2}$. Also, choosing $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$ yields a test procedure δ_4, which is defined by the values $c_1 = \mu_0 - 2.33\sigma n^{-1/2}$ and $c_2 = \mu_0 + 1.75\sigma n^{-1/2}$. The power functions $\pi(\mu|\delta_3)$ and $\pi(\mu|\delta_4)$ of these test procedures δ_3 and δ_4 are sketched

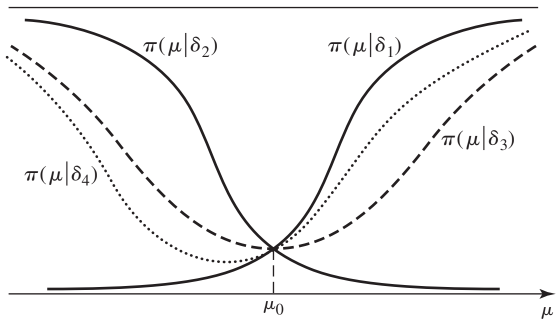


Figure 9.8 The power functions of four test procedures.

in Fig. 9.8, along with the power functions $\pi(\mu|\delta_1)$ and $\pi(\mu|\delta_2)$, which had previously been sketched in Figs. 9.6 and 9.7.

As the values of c_1 and c_2 in Eq. (9.4.2) or Eq. (9.4.3) are decreased, the power function $\pi(\mu|\delta)$ will become smaller for $\mu < \mu_0$ and larger for $\mu > \mu_0$. For $\alpha_0 = 0.05$, the limiting case is obtained by choosing $c_1 = -\infty$ and $c_2 = \mu_0 + 1.645\sigma n^{-1/2}$. The test procedure defined by these values is just δ_1. Similarly, as the values of c_1 and c_2 in Eq. (9.4.2) or Eq. (9.4.3) are increased, the power function $\pi(\mu|\delta)$ will become larger for $\mu < \mu_0$ and smaller for $\mu > \mu_0$. For $\alpha_0 = 0.05$, the limiting case is obtained by choosing $c_2 = \infty$ and $c_1 = \mu_0 - 1.645\sigma n^{-1/2}$. The test procedure defined by these values is just δ_2. Something between these two extreme limiting cases seems appropriate for hypotheses (9.4.1).

Selection of the Test Procedure

For a given sample size n, the values of the constants c_1 and c_2 in Eq. (9.4.2) should be chosen so that the size and shape of the power function are appropriate for the particular problem to be solved. In some problems, it is important not to reject the null hypothesis unless the data strongly indicate that μ differs greatly from μ_0. In such problems, a small value of α_0 should be used. In other problems, not rejecting the null hypothesis H_0 when μ is slightly larger than μ_0 is a more serious error than not rejecting H_0 when μ is slightly less than μ_0. Then it is better to select a test having a power function such as $\pi(\mu|\delta_4)$ in Fig. 9.8 than to select a test having a symmetric function such as $\pi(\mu|\delta_3)$.

In general, the choice of a particular test procedure in a given problem should be based both on the cost of rejecting H_0 when $\mu = \mu_0$ and on the cost, for each possible value of μ, of not rejecting H_0 when $\mu \neq \mu_0$. Also, when a test is being selected, the relative likelihoods of different values of μ should be considered. For example, if it is more likely that μ will be greater than μ_0 than that μ will be less than μ_0, then it is better to select a test for which the power function is large when $\mu > \mu_0$, and not so large when $\mu < \mu_0$, than to select one for which these relations are reversed.

Example 9.4.2

Egyptian Skulls. Suppose that, in Example 9.4.1, it is equally important to reject the null hypothesis that the mean breadth μ equals 140 when $\mu < 140$ as when $\mu > 140$. Then we should choose a test that rejects H_0 when the sample average $\overline{X}_n$ is either at most c_1 or at least c_2 where c_1 and c_2 are symmetric around 140. Suppose that we want a test of size $\alpha_0 = 0.05$. There are $n = 30$ skulls from 4000 B.C., so

$$c_1 = 140 - 1.96(26)^{1/2}30^{-1/2} = 138.18,$$

$$c_2 = 140 + 1.96(26)^{1/2}30^{-1/2} = 141.82.$$

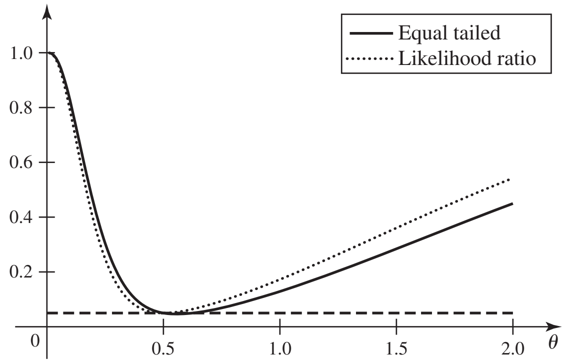


Figure 9.9 The power functions for the level $\alpha_0 = 0.05$ tests in Example 9.4.3 (equal tailed) and Example 9.4.4 (likelihood ratio). The horizontal line is at height 0.05.

The observed value of $\overline{X}_n$ is 131.37 in this case, and we would reject H_0 at the level of significance 0.05. ◂

In Examples 9.4.1 and 9.4.2, we would probably not wish to assume that the variance of the skull breadths was known to be 26, but rather we would assume that both the mean and the variance were unknown. We will see how to handle such a case in Sec. 9.5.

Other Distributions

The principles introduced above for samples from a normal distribution can be extended to any random sample. The details of implementation can be more tedious and less satisfying for other distributions.

Example 9.4.3 Service Times in a Queue. The manager in Example 9.3.2 models service times $X_1, \ldots, X_n$ as i.i.d. exponential random variables with parameter θ conditional on θ. Suppose that she wishes to test the null hypothesis $H_0 : \theta = 1/2$ versus the alternative $H_1 : \theta \neq 1/2$. For the one-sided alternative $\theta > 1/2$, we found (in Example 9.3.2) that the UMP level α_0 test was to reject H_0 if $T = \sum_{i=1}^n X_i$ is less than the α_0 quantile of the gamma distribution with parameters n and 1/2. By similar reasoning, the UMP level α_0 test of H_0 versus the other one-sided alternative $\theta < 1/2$ would be to reject H_0 if T is greater than the $1 - \alpha_0$ quantile of the gamma distribution with parameters n and 1/2. A simple way to construct a level α_0 test of $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$ would be to apply the same reasoning that we applied immediately after Eq. (9.4.2). That is, combine two one-sided tests with levels α_1 and α_2 where $\alpha_1 + \alpha_2 = \alpha_0$.

As a specific example, let $\alpha_1 = \alpha_2 = \alpha_0/2$, and let $G^{-1}(\cdot; n, 1/2)$ be the quantile function of the gamma distribution with parameters n and 1/2. Then, we reject H_0 if $T \leq G^{-1}(\alpha_0/2; n, 1/2)$ or $T \geq G^{-1}(1 - \alpha_0/2; n, 1/2)$. For the case of $\alpha_0 = 0.05$ and $n = 3$, the graph of the power function of this test appears in Fig. 9.9 together with the power function of the likelihood ratio test that will be derived in Example 9.4.4. ◂

An alternative test in Example 9.4.3 would be the likelihood ratio test. In Example 9.4.3, the likelihood ratio test requires solving some nonlinear equations.

Example 9.4.4 Service Times in a Queue. Instead of the ad hoc two-sided test constructed in Example 9.4.3, suppose that the manager decides to find a likelihood ratio test. Suppose that $\sum_{i=1}^{n} X_i = t$ is observed. The likelihood function is then

$$f_n(\boldsymbol{x}|\theta) = \theta^n \exp(-t\theta), \text{ for } \theta > 0.$$

The M.L.E. of θ is $\hat{\theta} = n/t$, so the likelihood ratio statistic from Definition 9.1.11 is

$$\Lambda(\boldsymbol{x}) = \frac{(1/2)^n \exp(-t/2)}{(n/t)^n \exp(-n)} = \left(\frac{t}{2n}\right)^n \exp(n - t/2). \tag{9.4.4}$$

The likelihood ratio test rejects H_0 if $\Lambda(\boldsymbol{x}) \le c$ for some constant c. From (9.4.4), we see that $\Lambda(\boldsymbol{x}) \le c$ is equivalent to $t \le c_1$ or $t \ge c_2$ where $c_1 < c_2$ satisfy

$$\left(\frac{c_1}{2n}\right)^n \exp(n - c_1/2) = \left(\frac{c_2}{2n}\right)^n \exp(n - c_2/2).$$

In order for the test to have level α_0, c_1 and c_2 must also satisfy

$$G(c_1; n, 1/2) + 1 - G(c_2; n, 1/2) = \alpha_0,$$

where $G(\cdot; n, 1/2)$ is the c.d.f. of the gamma distribution with parameters n and 1/2. Solving these two equations for c_1 and c_2 would give us the likelihood ratio test. Using numerical methods, the solution is $c_1 = 1.425$ and $c_2 = 15.897$. The power function of the likelihood ratio test is plotted in Fig. 9.9 together with the power function of the equal-tailed test. ◀

Composite Null Hypothesis

From one point of view, it makes little sense to carry out a test of the hypotheses (9.4.1) in which the null hypothesis H_0 specifies a single exact value μ_0 for the parameter μ. This is particularly true if we think of μ as the limit of the averages of increasing samples of future observations. Since it is inconceivable that μ will be *exactly* equal to μ_0 in any real problem, we know that the hypothesis H_0 cannot be true. Therefore, H_0 should be rejected as soon as it has been formulated.

This criticism is valid when it is interpreted literally. In many problems, however, the experimenter is interested in testing the null hypothesis H_0 that the value of μ is close to some specified value μ_0 against the alternative hypothesis that μ is not close to μ_0. In some of these problems, the simple hypothesis H_0 that $\mu = \mu_0$ can be used as an idealization or simplification for the purpose of choosing a decision. At other times, it is worthwhile to use a more realistic composite null hypothesis, which specifies that μ lies in an explicit interval around the value μ_0. We shall now consider hypotheses of this type.

Example 9.4.5

Testing an Interval Null Hypothesis. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance $\sigma^2 = 1$, and suppose that the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \ 9.9 \le \mu \le 10.1, \\ H_1&: \ \mu < 9.9 \text{ or } \mu > 10.1. \end{aligned} \tag{9.4.5}$$

Since the alternative hypothesis H_1 is two-sided, it is again appropriate to use a test procedure δ that rejects H_0 if either $\overline{X}_n \le c_1$ or $\overline{X}_n \ge c_2$. We shall determine the values of c_1 and c_2 for which the probability of rejecting H_0, when either $\mu = 9.9$ or $\mu = 10.1$, will be 0.05.

Let $\pi(\mu|\delta)$ denote the power function of δ. When $\mu = 9.9$, the random variable $n^{1/2}(\overline{X}_n - 9.9)$ has the standard normal distribution. Therefore,

$$\begin{aligned}\pi(9.9|\delta) &= \Pr(\text{Rejecting } H_0|\mu = 9.9)\\ &= \Pr(\overline{X}_n \le c_1|\mu = 9.9) + \Pr(\overline{X}_n \ge c_2|\mu = 9.9)\\ &= \Phi[n^{1/2}(c_1 - 9.9)] + 1 - \Phi[n^{1/2}(c_2 - 9.9)].\end{aligned} \tag{9.4.6}$$

Similarly, when $\mu = 10.1$, the random variable $n^{1/2}(\overline{X}_n - 10.1)$ has the standard normal distribution and

$$\pi(10.1|\delta) = \Phi[n^{1/2}(c_1 - 10.1)] + 1 - \Phi[n^{1/2}(c_2 - 10.1)]. \tag{9.4.7}$$

Both $\pi(9.9|\delta)$ and $\pi(10.1|\delta)$ must be made equal to 0.05. Because of the symmetry of the normal distribution, it follows that if the values of c_1 and c_2 are chosen symmetrically with respect to the value 10, then the power function $\pi(\mu|\delta)$ will be symmetric with respect to the point $\mu = 10$. In particular, it will then be true that $\pi(9.9|\delta) = \pi(10.1|\delta)$.

Accordingly, let $c_1 = 10 - c$ and $c_2 = 10 + c$. Then it follows from Eqs. (9.4.6) and (9.4.7) that

$$\pi(9.9|\delta) = \pi(10.1|\delta) = \Phi[n^{1/2}(0.1 - c)] + 1 - \Phi[n^{1/2}(0.1 + c)]. \tag{9.4.8}$$

The value of c must be chosen so that $\pi(9.9|\delta) = \pi(10.1|\delta) = 0.05$. Therefore, c must be chosen so that

$$\Phi[n^{1/2}(0.1 + c)] - \Phi[n^{1/2}(0.1 - c)] = 0.95. \tag{9.4.9}$$

For each given value of n, the value of c that satisfies Eq. (9.4.9) can be found by trial and error from a table of the standard normal distribution or using statistical software.

For example, if $n = 16$, then c must be chosen so that

$$\Phi(0.4 + 4c) - \Phi(0.4 - 4c) = 0.95. \tag{9.4.10}$$

After trying various values of c, we find that Eq. (9.4.10) will be satisfied when $c = 0.527$. Hence,

$$c_1 = 10 - 0.527 = 9.473 \text{ and } c_2 = 10 + 0.527 = 10.527.$$

Thus, when $n = 16$, the procedure δ rejects H_0 when either $\overline{X}_n \le 9.437$ or $\overline{X}_n \ge 10.527$. This procedure has a power function $\pi(\mu|\delta)$, which is symmetric with respect to the point $\mu = 10$ and for which $\pi(9.9|\delta) = \pi(10.1|\delta) = 0.05$. Furthermore, it is true that $\pi(\mu|\delta) < 0.05$ for $9.9 < \mu < 10.1$ and $\pi(\mu|\delta) > 0.05$ for $\mu < 9.9$ or $\mu > 10.1$. The function $\pi(\mu|\delta)$ is sketched in Fig. 9.10. ◀

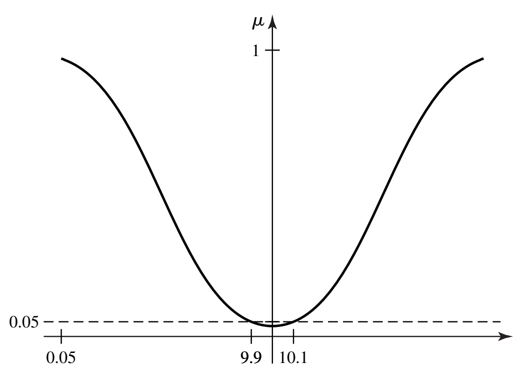


Figure 9.10 The power function $\pi(\mu|\delta)$ for a test of the hypotheses (9.4.5).

Unbiased Tests

Consider the general problem of testing the following hypotheses:

$$H_0: \quad \theta \in \Omega_0,$$
$$H_1: \quad \theta \in \Omega_1.$$

As usual, let $\pi(\theta|\delta)$ denote the power function of an arbitrary test procedure δ.

Definition 9.4.1 Unbiased Test. A test procedure δ is said to be *unbiased* if, for every $\theta \in \Omega_0$ and $\theta' \in \Omega_1$,

$$\pi(\theta|\delta) \le \pi(\theta'|\delta). \tag{9.4.11}$$

In words, δ is unbiased if its power function throughout Ω_1 is at least as large as it is throughout Ω_0.

If one closely examines Fig. 9.9, one sees that for values of θ slightly above 1/2, the power function of the equal-tailed test dips below 0.05 (the value of the power function at $\theta = 1/2$). This means that the test is not unbiased. This is typical in cases where the distribution of the test statistic T is not symmetric but a two-sided test is created by combining two one-sided tests. It is easy to see that an unbiased test would need to have a power function with derivative equal to 0 at $\theta = 1/2$; otherwise, it would dip below 0.05 on one side or the other of $\theta = 1/2$.

In many problems, the power function of every test is differentiable as a function of θ. In such cases, in order to create an unbiased level α_0 test δ of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$, we would need

$$\pi(\theta_0|\delta) = \alpha_0, \text{ and}$$
$$\left.\frac{d}{d\theta}\pi(\theta|\delta)\right|_{\theta=\theta_0} = 0. \tag{9.4.12}$$

Such equations would need to be solved numerically in any real problem. Typically, researchers don't think it is worth the trouble to solve such equations just to find an unbiased test.

Example 9.4.6 Service Times in a Queue. In Example 9.4.4, let $T = \sum_{i=1}^n X_i$. If we want an unbiased test of the form "reject H_0 if $T \le c_1$ or if $T \ge c_2$," the power function will be

$$\pi(\theta|\delta) = G(c_1; n, \theta) + 1 - G(c_2; n, \theta),$$

where $G(\cdot; n, \theta)$ is the c.d.f. of T given θ,

$$G(x; n, \theta) = \int_0^x \frac{\theta^n}{(n-1)!} t^{n-1} \exp(-t\theta)dt,$$

for $t > 0$. Eq. (9.4.12) requires that we compute the derivative of G with respect to θ. The derivative with respect to θ can be passed under the integral, and the result is

$$\begin{aligned}\frac{\partial}{\partial\theta}G(x; n, \theta) = &\int_0^x \frac{n\theta^{n-1}}{(n-1)!} t^{n-1} \exp(-t\theta)dt \\ &- \int_0^x t\frac{\theta^n}{(n-1)!} t^{n-1} \exp(-t\theta)dt.\end{aligned} \tag{9.4.13}$$

The reader can show (see Exercise 13 in this section) that (9.4.13) can be rewritten as

$$\frac{\partial}{\partial\theta}G(x; n, \theta) = \frac{n}{\theta}\left[G(x; n, \theta) - G(x; n+1, \theta)\right]. \tag{9.4.14}$$

For $\alpha_0 = 0.05$ and $n = 3$, the two equations we need to solve for c_1 and c_2 are

$$G(c_1; 3, 1/2) + 1 - G(c_2; 3, 1/2) = 0.05,$$
$$\frac{3}{1/2}\left[G(x; 3, 1/2) - G(x; 4, 1/2)\right] = 0.$$

Solving these two equations numerically gives the same solution as the likelihood ratio test to the number of significant digits reported in Example 9.4.4. This explains why the power function of the likelihood ratio test appears not to dip below 0.05 anywhere. ◀

Intuitively, the notion of an unbiased test sounds appealing. Since the goal of a test procedure is to reject H_0 when $\theta \in \Omega_1$ and not to reject H_0 when $\theta \in \Omega_0$, it seems desirable that the probability of rejecting H_0 should be at least as large when $\theta \in \Omega_1$ as it is whenever $\theta \in \Omega_0$. It can be seen that the test δ for which the power function is sketched in Fig. 9.10 is an unbiased test of the hypotheses (9.4.5). Also, among the four tests for which the power functions are sketched in Fig. 9.8, only δ_3 is an unbiased test of the hypotheses (9.4.1). Although it is beyond the scope of this book, one can show that δ_3 is UMP among all unbiased level $\alpha_0 = 0.05$ tests of (9.4.1).

The requirement that a test is to be unbiased can sometimes narrow the selection of a test procedure. However, unbiased procedures should be sought only under relatively special circumstances. For example, when testing the hypotheses (9.4.5), the statistician should use the unbiased test δ represented in Fig. 9.10 only under the following conditions: He believes that, for every value $a > 0$, it is just as important to reject H_0 when $\theta = 10.1 + a$ as to reject H_0 when $\theta = 9.9 - a$, and he also believes that these two values of θ are equally likely. In practice, the statistician might very well forego the use of an unbiased test in order to use a biased test that has higher power in certain regions of Ω_1 that he regards as particularly important or most likely to contain the true value of θ when H_0 is false.

In the remainder of this chapter, we shall consider special testing situations that arise very often in applied work. In these situations, there do not exist UMP tests. We shall study the most popular tests in these situations, and we shall show that these tests are likelihood ratio tests. However, in more advanced courses, it can be shown that the t tests and F tests derived in Sections 9.5, 9.6, and 9.7 are all UMP among various classes of unbiased tests of their sizes.

Summary

For the case of testing that the mean of a normal distribution with known variance equals a specific value against the two-sided alternative, one can construct level α_0 tests by combining the rejection regions of two one-sided tests of sizes α_1 and α_2 such that $\alpha_0 = \alpha_1 + \alpha_2$. A popular choice is $\alpha_1 = \alpha_2 = \alpha_0/2$. In this case, if $X_1, \ldots, X_n$ form a random sample from a normal distribution with mean μ and variance σ^2, one can test $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$ by rejecting H_0 if $\overline{X}_n > \mu_0 + \Phi^{-1}(1 - \alpha_0/2)\sigma/n^{1/2}$ or if $\overline{X}_n < \mu_0 - \Phi^{-1}(1 - \alpha_0/2)\sigma/n^{1/2}$, where Φ^{-1} is the quantile function of the standard normal distribution. A test is unbiased if its power function is greater at every point in the alternative hypothesis than at every point in the null hypothesis. The normal distribution test just described, with $\alpha_1 = \alpha_2 = \alpha_0/2$, is unbiased.

Exercises

1. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance 1, and it is desired to test the following hypotheses for a given number μ_0:

$$H_0: \quad \mu = \mu_0,$$
$$H_1: \quad \mu \neq \mu_0.$$

Consider a test procedure δ such that the hypothesis H_0 is rejected if either $\overline{X}_n \le c_1$ or $\overline{X}_n \ge c_2$, and let $\pi(\mu|\delta)$ denote the power function of δ. Determine the values of the constants c_1 and c_2 such that $\pi(\mu_0|\delta) = 0.10$ and the function $\pi(\mu|\delta)$ is symmetric with respect to the point $\mu = \mu_0$.

2. Consider again the conditions of Exercise 1, and suppose that

$$c_1 = \mu_0 - 1.96n^{-1/2}.$$

Determine the value of c_2 such that $\pi(\mu_0|\delta) = 0.10$.

3. Consider again the conditions of Exercise 1 and also the test procedure described in that exercise. Determine the smallest value of n for which $\pi(\mu_0|\delta) = 0.10$ and $\pi(\mu_0 + 1|\delta) = \pi(\mu_0 - 1|\delta) \ge 0.95$.

4. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance 1, and it is desired to test the following hypotheses:

$$H_0: \ 0.1 \le \mu \le 0.2,$$
$$H_1: \ \mu < 0.1 \text{ or } \mu > 0.2.$$

Consider a test procedure δ such that the hypothesis H_0 is rejected if either $\overline{X}_n \le c_1$ or $\overline{X}_n \ge c_2$, and let $\pi(\mu|\delta)$ denote the power function of δ. Suppose that the sample size is $n = 25$. Determine the values of the constants c_1 and c_2 such that $\pi(0.1|\delta) = \pi(0.2|\delta) = 0.07$.

5. Consider again the conditions of Exercise 4, and suppose also that $n = 25$. Determine the values of the constants c_1 and c_2 such that $\pi(0.1|\delta) = 0.02$ and $\pi(0.2|\delta) = 0.05$.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown, and it is desired to test the following hypotheses:

$$H_0: \quad \theta \le 3,$$
$$H_1: \quad \theta > 3.$$

- **a.** Show that for each level of significance α_0 $(0 \le \alpha_0 < 1)$, there exists a UMP test that specifies that H_0 should be rejected if $\max\{X_1, \ldots, X_n\} \ge c$.
- **b.** Determine the value of c for each possible value of α_0.

7. For a given sample size n and a given value of α_0, sketch the power function of the UMP test found in Exercise 6.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution described in Exercise 6, but suppose now that it is desired to test the following hypotheses:

$$H_0: \quad \theta \ge 3,$$
$$H_1: \quad \theta < 3.$$

- **a.** Show that at each level of significance α_0 $(0 < \alpha_0 < 1)$, there exists a UMP test that specifies that H_0 should be rejected if $\max\{X_1, \ldots, X_n\} \le c$.
- **b.** Determine the value of c for each possible value of α_0.

9. For a given sample size n and a given value of α_0, sketch the power function of the UMP test found in Exercise 8.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution described in Exercise 6, but suppose now that it is desired to test the following hypotheses:

$$\begin{aligned} H_0: &\quad \theta = 3, \\ H_1: &\quad \theta \neq 3. \end{aligned} \tag{9.4.15}$$

Consider a test procedure δ such that the hypothesis H_0 is rejected if either $\max\{X_1, \ldots, X_n\} \le c_1$ or $\max\{X_1, \ldots, X_n\} \ge c_2$, and let $\pi(\theta|\delta)$ denote the power function of δ.

- **a.** Determine the values of the constants c_1 and c_2 such that $\pi(3|\delta) = 0.05$ and δ is unbiased.
- **b.** Prove that the test found in part (a) is UMP of level 0.05 for testing the hypotheses in (9.4.15). *Hint:* Compare this test to the UMP tests of level $\alpha_0 = 0.05$ in Exercises 6 and 8.
- **c.** Determine the values of the constants c_1 and c_2 such that $\pi(3|\delta) = 0.05$ and δ is unbiased.

11. Consider again the conditions of Exercise 1. Determine the values of the constants c_1 and c_2 such that $\pi(\mu_0|\delta) = 0.10$ and δ is unbiased.

12. Let X have the exponential distribution with parameter β. Suppose that we wish to test the hypotheses

$$H_0: \quad \beta = 1,$$
$$H_1: \quad \beta \neq 1.$$

We shall use a test procedure that rejects H_0 if either $X \le c_1$ or $X \ge c_2$.

- **a.** Find the equation that must be satisfied by c_1 and c_2 in order for the test procedure to have level of significance α_0.
- **b.** Find a pair of finite, nonzero values (c_1, c_2) such that the test procedure has level of significance $\alpha_0 = 0.1$.

13. Prove Eq. (9.4.14) in Example 9.4.6.*Hint:* Both parts of the integrand in Eq. (9.4.13) differ from gamma distribution p.d.f.'s by some factor that does not depend on t.

9.5 The t Test

We begin the treatment of several special cases of testing hypotheses about parameters of a normal distribution. In this section, we handle the case in which both the mean and the variance are unknown. We develop tests for hypotheses concerning the mean. These tests will be based on the t distribution.

Testing Hypotheses about the Mean of a Normal Distribution When the Variance Is Unknown

Example 9.5.1 Nursing Homes in New Mexico. In Example 8.6.3, we described a study of medical in-patient days in nursing homes in New Mexico. As in that example, we shall model the numbers of medical in-patient days as a random sample of $n = 18$ normal random variables with unknown mean μ and unknown variance σ^2. Suppose that we are interested in testing the hypotheses $H_0 : \mu \geq 200$ versus $H_1 : \mu < 200$. What test should we use, and what are its properties? ◄

In this section we shall consider the problem of testing hypotheses about the mean of a normal distribution when both the mean and the variance are unknown. Specifically, we shall suppose that the random variables $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ and the variance σ^2 are unknown, and we shall consider testing the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu \leq \mu_0, \\ H_1&: \quad \mu > \mu_0. \end{aligned} \tag{9.5.1}$$

The parameter space Ω in this problem comprises every two-dimensional vector (μ, σ^2), where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. The null hypothesis H_0 specifies that the vector (μ, σ^2) lies in the subset Ω_0 of Ω, comprising all vectors for which $\mu \leq \mu_0$ and $\sigma^2 > 0$, as illustrated in Fig. 9.11. The alternative hypothesis H_1 specifies that (μ, σ^2) belongs to the subset Ω_1 of Ω, comprising all the vectors that do not belong to Ω_0.

In Example 9.1.17 on page 543, we showed how to derive a test of the hypotheses (9.5.1) from a one-sided confidence interval for μ. To be specific, define $\overline{X}_n = \sum_{i=1}^{n} X_i/n$, $\sigma' = (\sum_{i=1}^{n}(X_i - \overline{X}_n)^2/[n-1])^{1/2}$, and

$$U = n^{1/2}\frac{\overline{X}_n - \mu_0}{\sigma'}. \tag{9.5.2}$$

The test rejects H_0 if $U \geq c$. When $\mu = \mu_0$, it follows from Theorem 8.4.2 that the distribution of the statistic U defined in Eq. (9.5.2) is the t distribution with $n - 1$

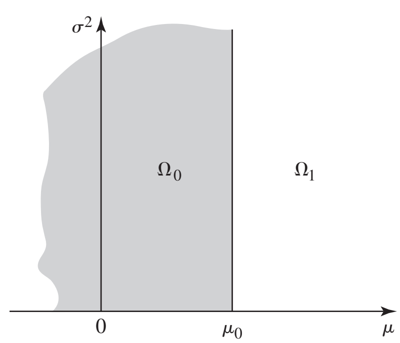


Figure 9.11 The subsets Ω_0 and Ω_1 of the parameter space Ω for the hypotheses (9.5.1).

degrees of freedom, regardless of the value of σ^2. For this reason, tests based on U are called *t tests*. When we want to test

$$\begin{aligned} H_0&: \quad \mu \geq \mu_0, \\ H_1&: \quad \mu < \mu_0, \end{aligned} \tag{9.5.3}$$

the test is of the form "reject H_0 if $U \leq c$."

Example 9.5.2 Nursing Homes in New Mexico. In Example 9.5.1, if we desired a level α_0 test, we could use the t test that rejects H_0 if the statistic U in Eq. (9.5.2) is at most equal to the constant c chosen to make the size of the test equal to α_0. ◀

Properties of the t Tests

Theorem 9.5.1 gives some useful properties of t tests.

Theorem 9.5.1 Level and Unbiasedness of t Tests. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the normal distribution with mean μ and variance σ^2, let U be the statistic in Eq. (9.5.2), and let c be the $1 - \alpha_0$ quantile of the t distribution with $n - 1$ degrees of freedom. Let δ be the test that rejects H_0 in (9.5.1) if $U \geq c$. The power function $\pi(\mu, \sigma^2|\delta)$ has the following properties:

i. $\pi(\mu, \sigma^2|\delta) = \alpha_0$ when $\mu = \mu_0$,
ii. $\pi(\mu, \sigma^2|\delta) < \alpha_0$ when $\mu < \mu_0$,
iii. $\pi(\mu, \sigma^2|\delta) > \alpha_0$ when $\mu > \mu_0$,
iv. $\pi(\mu, \sigma^2|\delta) \to 0$ as $\mu \to -\infty$,
v. $\pi(\mu, \sigma^2|\delta) \to 1$ as $\mu \to \infty$.

Furthermore, the test δ has size α_0 and is unbiased.

Proof If $\mu = \mu_0$, then U has the t distribution with $n - 1$ degrees of freedom. Hence,

$$\pi(\mu_0, \sigma^2|\delta) = \Pr(U \geq c|\mu 0, \sigma^2) = \alpha_0.$$

This proves (i) above. For (ii) and (iii), define

$$U^* = \frac{n^{1/2}(\overline{X}_n - \mu)}{\sigma'} \quad \text{and} \quad W = \frac{n^{1/2}(\mu_0 - \mu)}{\sigma'}.$$

Then $U = U^* - W$. First, assume that $\mu < \mu_0$ so that $W > 0$. It follows that

$$\begin{aligned} \pi(\mu, \sigma^2|\delta) &= \Pr(U \geq c|\mu, \sigma^2) = \Pr(U^* - W \geq c|\mu, \sigma^2) \\ &= \Pr(U^* \geq c + W|\mu, \sigma^2) < \Pr(U^* \geq c|\mu, \sigma^2). \end{aligned} \tag{9.5.4}$$

Since U^* has the t distribution with $n - 1$ degrees of freedom, the last probability in (9.5.4) is α_0. This proves (ii). For (iii), let $\mu > \mu_0$ so that $W < 0$. The less-than in (9.5.4) becomes a greater-than, and (iii) is proven.

That the size of the test is α_0 is immediate from parts (i) and (ii). That the test is unbiased is immediate from parts (i) and (iii).

The proofs of (iv) and (v) are more difficult and will not be given here in detail. Intuitively, if μ is very large, then W in Eq. (9.5.4) will tend to be very negative, and the probability will be close to 1 that $U^* \geq c + W$. Similarly, if μ is very much less than 0, then W will tend to be very positive, and the chance of $U^* \geq c + W$ will be close to 0. ■

For the hypotheses of Eq. (9.5.3), very similar properties hold.

Corollary 9.5.1 *t* Tests for Hypotheses of Eq. (9.5.3). Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the normal distribution with mean μ and variance σ^2, let U be the statistic in Eq. (9.5.2), and let c be the α_0 quantile of the t distribution with $n-1$ degrees of freedom. Let δ be the test that rejects H_0 in (9.5.3) if $U \le c$. The power function $\pi(\mu, \sigma^2|\delta)$ has the following properties:

i. $\pi(\mu, \sigma^2|\delta) = \alpha_0$ when $\mu = \mu_0$,
ii. $\pi(\mu, \sigma^2|\delta) > \alpha_0$ when $\mu < \mu_0$,
iii. $\pi(\mu, \sigma^2|\delta) < \alpha_0$ when $\mu > \mu_0$,
iv. $\pi(\mu, \sigma^2|\delta) \to 1$ as $\mu \to -\infty$,
v. $\pi(\mu, \sigma^2|\delta) \to 0$ as $\mu \to \infty$.

Furthermore, the test δ has size α_0 and is unbiased.

Example 9.5.3 Nursing Homes in New Mexico. In Examples 9.5.1 and 9.5.2, suppose that we desire a test with level of significance $\alpha_0 = 0.1$. Then we reject H_0 if $U \le c$ where c is the 0.1 quantile of the t distribution with 17 degrees of freedom, namely, -1.333. Using the data from Example 8.6.3, we calculate the observed value of $\overline{X}_{18} = 182.17$ and $\sigma' = 72.22$. The observed value of U is then $(17)^{1/2}(182.17 - 200)/72.22 = -1.018$. We would not reject $H_0 : \mu \ge 200$ at level of significance 0.1, because the observed value of U is greater than -1.333. ◀

***p*-Values for *t* Tests** The p-value from the observed data and a specific test is the smallest α_0 such that we would reject the null hypothesis at level of significance α_0. For the t tests that we have just discussed, it is straightforward to compute the p-values.

Theorem 9.5.2 *p*-Values for *t* Tests. Suppose that we are testing either the hypotheses in Eq. (9.5.1) or the hypotheses in Eq. (9.5.3). Let u be the observed value of the statistic U in Eq. (9.5.2), and let $T_{n-1}(\cdot)$ be the c.d.f. of the t distribution with $n-1$ degrees of freedom. Then the p-value for the hypotheses in Eq. (9.5.1) is $1 - T_{n-1}(u)$ and the p-value for the hypotheses in Eq. (9.5.3) is $T_{n-1}(u)$.

Proof Let $T_{n-1}^{-1}(\cdot)$ stand for the quantile function of the t distribution with $n-1$ degrees of freedom. This is the inverse of the strictly increasing function T_{n-1}. We would reject the hypotheses in Eq. (9.5.1) at level α_0 if and only if $u \ge T_{n-1}^{-1}(1-\alpha_0)$, which is equivalent to $T_{n-1}(u) \ge 1 - \alpha_0$, which is equivalent to $\alpha_0 \ge 1 - T_{n-1}(u)$. Hence, the smallest level α_0 at which we could reject H_0 is $1 - T_{n-1}(u)$. Similarly, we would reject the hypotheses in Eq. (9.5.3) if and only if $u \le T_{n-1}^{-1}(\alpha_0)$, which is equivalent to $\alpha_0 \ge T_{n-1}(u)$. ■

Example 9.5.4 Lengths of Fibers. Suppose that the lengths in millimeters of metal fibers produced by a certain process have the normal distribution with unknown mean μ and unknown variance σ^2, and the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \mu \le 5.2, \\ H_1&: \quad \mu > 5.2. \end{aligned} \tag{9.5.5}$$

Suppose that the lengths of 15 fibers selected at random are measured, and it is found that the sample mean $\overline{X}_{15}$ is 5.4 and $\sigma' = 0.4226$. Based on these measurements, we shall carry out a t test at the level of significance $\alpha_0 = 0.05$.

Since $n = 15$ and $\mu_0 = 5.2$, the statistic U defined by Eq. (9.5.2) will have the t distribution with 14 degrees of freedom when $\mu = 5.2$. It is found in the table of the

t distribution that $T_{14}^{-1}(0.95) = 1.761$. Hence, the null hypothesis H_0 will be rejected if $U > 1.761$. Since the numerical value of U calculated from Eq. (9.5.2) is 1.833, H_0 would be rejected at level 0.05.

With observed value $u = 1.833$ for the statistic U and $n = 15$, we can compute the p-value for the hypotheses (9.5.1) using computer software that includes the c.d.f. of various t distributions. In particular, we find $1 - T_{14}(1.833) = 0.0441$. ◄

The Complete Power Function For all values of μ, the power function of a t test can be determined if we know the distribution of U defined in Eq. (9.5.2). We can rewrite U as

$$U = \frac{n^{1/2}(\overline{X}_n - \mu_0)/\sigma}{\sigma'/\sigma}. \tag{9.5.6}$$

The numerator of the right side in Eq. (9.5.6) has the normal distribution with mean $n^{1/2}(\mu - \mu_0)/\sigma$ and variance 1. The denominator is the square-root of a χ^2 random variable divided by its degrees of freedom, $n - 1$. Were it not for the nonzero mean, the ratio would have the t distribution with $n - 1$ degrees of freedom as we have already shown. When the mean of the numerator is not 0, U has a *noncentral t distribution*.

Definition 9.5.1 Noncentral t Distributions. Let Y and W be independent random variables with W having the normal distribution with mean ψ and variance 1 and Y having the χ^2 distribution with m degrees of freedom. Then the distribution of

$$X = \frac{W}{\left(\frac{Y}{m}\right)^{1/2}},$$

is called the *noncentral t distribution with m degrees of freedom and noncentrality parameter* ψ. We shall let $T_m(t|\psi)$ denote the c.d.f. of this distribution. That is, $T_m(t|\psi) = \Pr(X \le t)$.

It should be obvious that the noncentral t distribution with m degrees of freedom and noncentrality parameter $\psi = 0$ is also the t distribution with m degrees of freedom. The following result is also immediate from Definition 9.5.1.

Theorem 9.5.3 Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. The distribution of the statistic U in Eq. (9.5.2) is the noncentral t distribution with $n - 1$ degrees of freedom and noncentrality parameter $\psi = n^{1/2}(\mu - \mu_0)/\sigma$. Let δ be the test that rejects $H_0 : \mu \le \mu_0$ when $U \ge c$. Then the power function of δ is $\pi(\mu, \sigma^2|\delta) = 1 - T_{n-1}(c|\psi)$. Let δ' be the test that rejects $H_0 : \mu \ge \mu_0$ when $U \le c$. Then the power function of δ' is $\pi(\mu, \sigma^2|\delta') = T_{n-1}(c|\psi)$. ■

In Exercise 11, you can prove that $1 - T_m(t|\psi) = T_m(-t| - \psi)$. There are computer programs to calculate the c.d.f.'s of noncentral t distributions, and some statistical software packages include such programs. Figure 9.12 plots the power functions of level 0.05 and level 0.01 t tests for various degrees of freedom and various values of the noncentrality parameter. The horizontal axis is labeled $|\psi|$ because the same graphs can be used for both types of one-sided hypotheses. The next example illustrates how to use Fig. 9.12 to approximate the power function.

Example 9.5.5 Lengths of Fibers. In Example 9.5.4, we tested the hypotheses (9.5.5) at level 0.05. Suppose that we are interested in the power of our test when μ is not equal to 5.2. In

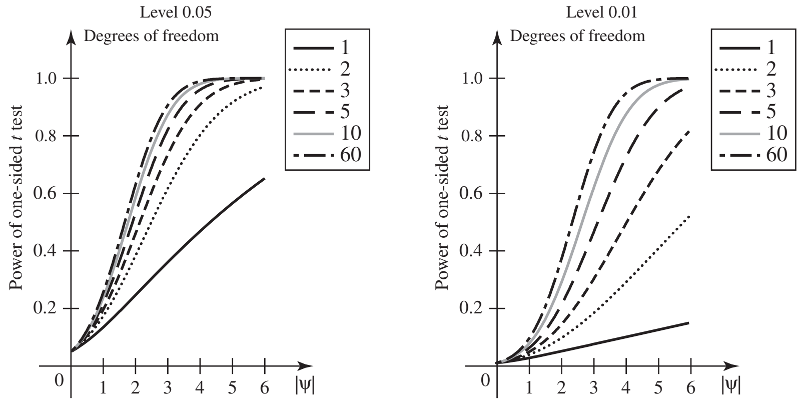


Figure 9.12 The power functions on the alternative of one-sided level 0.05 and level 0.01 t tests with various degrees of freedom for various values of the noncentrality parameter ψ.

particular, suppose that we are interested in the power when $\mu = 5.2 + \sigma/2$, one-half standard deviation above 5.2. Then the noncentrality parameter is

$$\psi = 15^{1/2}\left(\frac{5.2 + \sigma/2 - 5.2}{\sigma}\right) = 1.936.$$

There is no curve for 14 degrees of freedom in Fig. 9.12; however, there is not much difference between the curves for 10 and 60 degrees of freedom, so we can assume that our answer is somewhere between those two. If we look at the level 0.05 plot in Fig. 9.12 and move up from 1.936 (about 2) on the horizontal axis until we get a little above the curve for degrees of freedom equal to 10, we find that the power is about 0.6. (The actual power is 0.578.) ◀

Note: Power is a Function of the Noncentrality Parameter. In Example 9.5.5, we cannot answer a question like "What is the power of a level 0.05 test when $\mu = 5.5$?" The reason is that the power is a function of both μ and σ through the noncentrality parameter. (See Exercise 6.) For each possible σ and $\mu = 5.5$, the noncentrality parameter is $\psi = 15^{1/2} \times 0.3/\sigma$, which varies from 0 to ∞ depending on σ. This is why, whenever we want a numerical value for the power of a t test, we need either to specify both μ and σ or to specify how far μ is from μ_0 in multiples of σ.

Choosing a Sample Size It is possible to use the power function of a test to help determine what would be an appropriate sample size to observe.

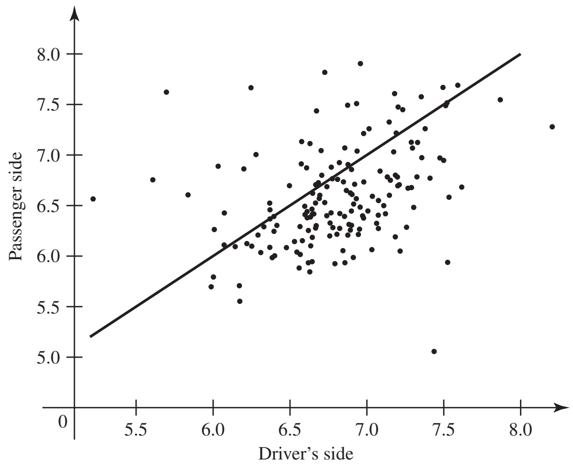
Example 9.5.6

Lengths of Fibers. In Example 9.5.5, we found that the power of the test was 0.578 when $\mu = 5.2 + \sigma/2$. Suppose that we want the power to be close to 0.8, when $\mu = 5.2 + \sigma/2$. It will take more than $n = 15$ observations to achieve this. In Fig. 9.12, we can see what size of noncentrality parameter ψ that we need in order for the power to reach 0.8. For degrees of freedom between 10 and 60, we need ψ to be about 2.5. But $\psi = n^{1/2}/2$ when $\mu = 5.2 + \sigma/2$. So we need $n = 25$ approximately. Precise calculation shows that, with $n = 25$, the power of the level 0.05 test is 0.7834 when $\mu = 5.2 + \sigma/2$. With $n = 26$, the power is 0.7981, and with $n = 27$ the power is 0.8118. ◀

The Paired t Test

In many experiments, the same variable is measured under two different conditions on the same experimental unit, and we are interested in whether the mean value is

Figure 9.13 Plot of logarithms of head injury measures for dummies on driver's side and passenger's side. The line indicates where the two measures are equal.



greater in one condition than in the other. In such cases, it is common to subtract the two measurements and treat the differences as a random sample from a normal distribution. We can then test hypotheses concerning the mean of the differences.

Example 9.5.7

Crash Test Dummies. The National Transportation Safety Board collects data from crash tests concerning the amount and location of damage on dummies placed in the tested cars. In one series of tests, one dummy was placed in the driver's seat and another was placed in the front passenger's seat of each car. One variable measured was the amount of injury to the head for each dummy. Figure 9.13 shows a plot of the pairs of logarithms of head injury measures for dummies in the two different seats. Among other things, interest lies in whether and/or to what extent the amount of head injury differs between the driver's seat and the passenger's seat. Let $X_1, \dots, X_n$ be the differences between the logarithms of head injury measures for driver's side and passenger's side. We can model $X_1, \dots, X_n$ as a random sample from a normal distribution with mean μ and variance σ^2. Suppose that we wish to test the null hypothesis $H_0 : \mu \le 0$ against the alternative $H_1 : \mu > 0$ at level $\alpha_0 = 0.01$. There are $n = 164$ cars represented in Fig. 9.13. The test would be to reject H_0 if $U \ge T_{163}^{-1}(0.99) = 2.35$.

The average of the differences of the coordinates in Fig. 9.13 is $\overline{x}_n = 0.2199$. The value of σ' is 0.5342. The statistic U is then 5.271. This is larger than 2.35, and the null hypothesis would be rejected at level 0.01. Indeed, the p-value is less than 1.0×10^{-6}.

Suppose also that we are interested in the power function under H_1 of the level 0.01 test. Suppose that the mean difference between driver's side and passenger's side logarithm of head injury is $\sigma/4$. Then the noncentrality parameter is $(164)^{1/2}/4 = 3.20$. In the right panel of Fig. 9.12, it appears that the power is just about 0.8. (In fact, it is 0.802.) ◄

Testing with a Two-Sided Alternative

Example 9.5.8

Egyptian Skulls. In Examples 9.4.1 and 9.4.2, we modeled the breadths of skulls from 4000 B.C. as a random sample of size $n = 30$ from a normal distribution with unknown mean μ and known variance. We shall now generalize that model to allow the more realisitc assumption that the variance σ^2 is unknown. Suppose that we wish to test the null hypothesis $H_0 : \mu = 140$ versus the alternative hypothesis $H_1 : \mu \ne 140$. We can still calculate the statistic U in Eq. (9.5.2), but now it would make sense to reject

H_0 if either $U \le c_1$ or $U \ge c_2$ for suitably chosen numbers c_1 and c_2. How should we choose c_1 and c_2, and what are the properties of the resulting test? ◀

As before, assume that $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown. Suppose now that the following hypotheses are to be tested:

$$\begin{aligned} H_0&:\quad \mu = \mu_0,\\ H_1&:\quad \mu \ne \mu_0. \end{aligned} \tag{9.5.7}$$

Here, the alternative hypothesis H_1 is two-sided.

In Example 9.1.15, we derived a level α_0 test of the hypotheses (9.5.7) from the confidence interval that was developed in Sec. 8.5. That test has the form "reject H_0 if $|U| \ge T_{n-1}^{-1}(1-\alpha_0/2)$," where T_{n-1}^{-1} is the quantile function of the t distribution with $n-1$ degrees of freedom and U is defined in Eq. (9.5.2).

Example 9.5.9 Egyptian Skulls. In Example 9.5.8, suppose that we want a level $\alpha_0 = 0.05$ test of $H_0: \mu = 140$ versus $H_1: \mu \ne 140$. If we use the test described above (derived in Example 9.1.15), then the two numbers c_1 and c_2 will be of opposite signs and equal in magnitude. Specifically, $c_1 = -T_{29}^{-1}(0.975) = -2.045$ and $c_2 = 2.045$. The observed value of $\overline{X}_{30}$ is 131.37, and the observed value of σ' is 5.129. The observed value u of the statistic U is $u = (30)^{1/2}(131.37 - 140)/5.129 = -9.219$. This is less than -2.045, so we would reject H_0 at level 0.05. ◀

Example 9.5.10 Lengths of Fibers. We shall consider again the problem discussed in Example 9.5.4, but we shall suppose now that, instead of the hypotheses (9.5.5), the following hypotheses are to be tested:

$$\begin{aligned} H_0&:\quad \mu = 5.2,\\ H_1&:\quad \mu \ne 5.2. \end{aligned} \tag{9.5.8}$$

We shall again assume that the lengths of 15 fibers are measured, and the value of U calculated from the observed values is 1.833. We shall test the hypotheses (9.5.8) at the level of significance $\alpha_0 = 0.05$.

Since $\alpha_0 = 0.05$, our critical value will be the $1 - 0.05/2 = 0.975$ quantile of the t distribution with 14 degrees of freedom. From the table of t distributions in this book, we find $T_{14}^{-1}(0.975) = 2.145$. So the t test specifies rejecting H_0 if either $U \le -2.145$ or $U \ge 2.145$. Since $U = 1.833$, the hypothesis H_0 would not be rejected. ◀

The numerical values in Examples 9.5.4 and 9.5.10 emphasize the importance of deciding whether the appropriate alternative hypothesis in a given problem is one-sided or two-sided. When the hypotheses (9.5.5) were tested at the level of significance 0.05, the hypothesis H_0 that $\mu \le 5.2$ was rejected. When the hypotheses (9.5.8) were tested at the same level of significance, and the same data were used, the hypothesis H_0 that $\mu = 5.2$ was not rejected.

Power Functions of Two-Sided Tests The power function of the test δ that rejects $H_0: \mu = \mu_0$ when $|U| \ge c$, where $c = T_{n-1}^{-1}(1-\alpha_0/2)$, can be found by using the noncentral t distribution. If $\mu \ne \mu_0$, then U has the noncentral t distribution with $n-1$ degrees of freedom and noncentrality parameter $\psi = n^{1/2}(\mu - \mu_0)/\sigma$, just as it did when we tested one-sided hypotheses. The power function of δ is then

$$\pi(\mu, \sigma^2|\delta) = T_{n-1}(-c|\psi) + 1 - T_{n-1}(c|\psi).$$

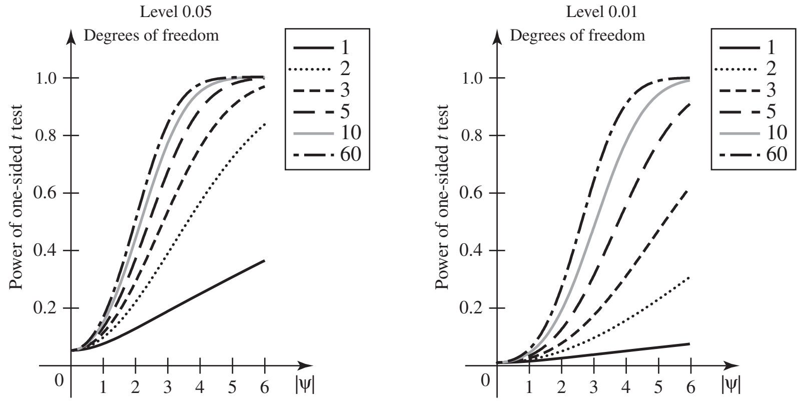


Figure 9.14 The power functions of two-sided level 0.05 and level 0.01 t tests with various degrees of freedom for various values of the noncentrality parameter ψ.

Figure 9.14 plots these power functions for various degrees of freedom and noncentrality parameters. We could use Fig. 9.14 to find the power of the test in Example 9.5.10 when $\mu = 5.2 + \sigma/2$, that is, when $\psi = 1.936$. It appears to be about 0.45. (The actual power is 0.438.)

Theorem 9.5.4 ***p*-Values for Two-Sided *t* Tests.** Suppose that we are testing the hypotheses in Eq. (9.5.7). Let u be the observed value of the statistic U, and let $T_{n-1}(\cdot)$ be the c.d.f. of the t distribution with $n-1$ degrees of freedom. Then the p-value is $2[1 - T_{n-1}(|u|)]$.

Proof Let $T_{n-1}^{-1}(\cdot)$ stand for the quantile function of the t distribution with $n-1$ degrees of freedom. We would reject the hypotheses in Eq. (9.5.7) at level α_0 if and only if $|u| \ge T_{n-1}^{-1}(1 - \alpha_0/2)$, which is equivalent to $T_{n-1}(|u|) \ge 1 - \alpha_0/2$, which is equivalent to $\alpha_0 \ge 2[1 - T_{n-1}(|u|)]$. Hence, the smallest level α_0 at which we could reject H_0 is $2[1 - T_{n-1}(|u|)]$. ■

Example 9.5.11 **Lengths of Fibers.** In Example 9.5.10, the p-value is $2[1 - T_{14}(1.833)] = 0.0882$. Note that this is twice the p-value when the hypotheses were (9.5.1). ◂

For t tests, if the p-value for testing hypotheses (9.5.1) or (9.5.3) is p, then the p-value for hypotheses (9.5.7) is the smaller of $2p$ and $2(1-p)$.

The t Test as a Likelihood Ratio Test

We introduced likelihood ratio tests in Sec. 9.1. We can compute such tests for the hypotheses of this section.

Example 9.5.12 **Likelihood Ratio Test of One-Sided Hypotheses about the Mean of a Normal Distribution.** Consider the hypotheses (9.5.1). After the values $x_1, \ldots, x_n$ in the random sample have been observed, the likelihood function is

$$f_n(\boldsymbol{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]. \tag{9.5.9}$$

In this case, $\Omega_0 = \{(\mu, \sigma^2) : \mu \le \mu_0\}$ and $\Omega_1 = \{(\mu, \sigma^2) : \mu > \mu_0\}$. The likelihood ratio statistic is

$$\Lambda(\boldsymbol{x}) = \frac{\sup_{\{(\mu,\sigma^2):\mu>\mu_0\}} f_n(\boldsymbol{x}|\mu,\sigma^2)}{\sup_{(\mu,\sigma^2)} f_n(\boldsymbol{x}|\mu,\sigma^2)}. \tag{9.5.10}$$

We shall now derive an explicit form for the likelihood ratio test based on (9.5.10). As in Sec. 7.5, we shall let $\hat{\mu}$ and $\hat{\sigma}^2$ denote the M.L.E.'s of μ and σ^2 when it is known only that the point (μ, σ^2) belongs to the parameter space Ω. It was shown in Example 7.5.6 that

$$\hat{\mu} = \overline{x}_n \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

It follows that the denominator of $\Lambda(\boldsymbol{x})$ equals

$$\sup_{(\mu,\sigma^2)} f_n(\boldsymbol{x}|\mu,\sigma^2) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}}\exp\left(-\frac{n}{2}\right). \tag{9.5.11}$$

Similarly, we shall let $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ denote the M.L.E.'s of μ and σ^2 when the point (μ, σ^2) is constrained to lie in the subset Ω_0. Suppose first that the observed sample values are such that $\overline{x}_n \le \mu_0$. Then the point $(\hat{\mu}, \hat{\sigma}^2)$ will lie in Ω_0 so that $\hat{\mu}_0 = \hat{\mu}$ and $\hat{\sigma}_0^2 = \hat{\sigma}^2$ and the numerator of $\Lambda(\boldsymbol{x})$ also equals (9.5.11). In this case, $\Lambda(\boldsymbol{x}) = 1$.

Next, suppose that the observed sample values are such that $\overline{x}_n > \mu_0$. Then the point $(\hat{\mu}, \hat{\sigma}^2)$ does not lie in Ω_0. In this case, it can be shown that $f_n(\boldsymbol{x}|\mu,\sigma^2)$ attains its maximum value among all points $(\mu, \sigma^2) \in \Omega_0$ if μ is chosen to be as close as possible to $\overline{x}_n$. The value of μ closest to $\overline{x}_n$ among all points in the subset Ω_0 is $\mu = \mu_0$. Hence, $\hat{\mu}_0 = \mu_0$. In turn, it can be shown, as in Example 7.5.6, that the M.L.E. of σ^2 will be

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_0)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2.$$

In this case, the numerator of $\Lambda(\boldsymbol{x})$ is then

$$\sup_{\{(\mu,\sigma^2):\mu>\mu_0\}} f_n(\boldsymbol{x}|\mu,\sigma^2) = \frac{1}{(2\pi\hat{\sigma}_0^2)^{n/2}}\exp\left(-\frac{n}{2}\right). \tag{9.5.12}$$

Taking the ratio of (9.5.12) to (9.5.11), we find that

$$\Lambda(\boldsymbol{x}) = \begin{cases} \left(\dfrac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} & \text{if } \overline{x}_n > \mu_0, \\ 1 & \text{otherwise.} \end{cases} \tag{9.5.13}$$

Next, use the relation

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + n(\overline{x}_n - \mu_0)^2$$

to write the top branch of (9.5.13) as

$$\left[1 + \frac{n(\overline{x}_n - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}\right]^{-n/2}. \tag{9.5.14}$$

If u is the observed value of the statistic U in Eq. (9.5.2), then one can easily check that

$$\frac{n(\overline{x}_n - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2} = \frac{u^2}{n-1}.$$

It follows that $\Lambda(\boldsymbol{x})$ is a nonincreasing function of u. Hence, for $k < 1$, $\Lambda(\boldsymbol{x}) \le k$ if and only if $u \ge c$, where

$$c = \left(\left[\frac{1}{k^{2/n}} - 1\right](n-1)\right)^{1/2}.$$

It follows that the likelihood ratio test is a t test. ◄

It is not difficult to adapt the argument in Example 9.5.12 to find the likelihood ratio tests for hypotheses (9.5.3) and (9.5.7). (See Exercises 17 and 18, for example.)

Summary

When $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2, we can test hypotheses about μ by using the fact that $n^{1/2}(\overline{X}_n - \mu)/\sigma'$ has the t distribution with $n-1$ degrees of freedom. Let T_{n-1}^{-1} denote the quantile function of the t distribution with $n-1$ degrees of freedom. Then, to test $H_0: \mu \le \mu_0$ versus $H_1: \mu > 0$ at level α_0, for instance, we reject H_0 if $n^{1/2}(\overline{X}_n - \mu_0)/\sigma' > T_{n-1}^{-1}(1-\alpha_0)$. To test $H_0: \mu = \mu_0$ versus $H_1: \mu \ne \mu_0$, reject H_0 if $|n^{1/2}(\overline{X}_n - \mu_0)/\sigma'| \ge T_{n-1}^{-1}(1-\alpha_0/2)$. The power functions of each of these tests can be written in terms of the c.d.f. of a noncentral t distribution with $n-1$ degrees of freedom and noncentrality parameter $\psi = n^{1/2}(\mu - \mu_0)/\sigma$.

Exercises

1. Use the data in Example 8.5.4, comprising a sample of $n = 10$ lactic acid measurements in cheese. Assume, as we did there, that the lactic acid measurements are a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Suppose that we wish to test the following hypotheses:

$$H_0: \quad \mu \le 1.2,$$
$$H_1: \quad \mu > 1.2.$$

a. Perform the level $\alpha_0 = 0.05$ test of these hypotheses.

b. Compute the p-value.

2. Suppose that nine observations are selected at random from the normal distribution with unknown mean μ and unknown variance σ^2, and for these nine observations it is found that $\overline{X}_n = 22$ and $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 72$.

a. Carry out a test of the following hypotheses at the level of significance 0.05:

$$H_0: \quad \mu \le 20,$$
$$H_1: \quad \mu > 20.$$

b. Carry out a test of the following hypotheses at the level of significance 0.05 by using the two-sided t test:

$$H_0: \quad \mu = 20,$$
$$H_1: \quad \mu \ne 20.$$

c. From the data, construct the observed confidence interval for μ with confidence coefficient 0.95.

3. The manufacturer of a certain type of automobile claims that under typical urban driving conditions the automobile will travel on average at least 20 miles per gallon of gasoline. The owner of this type of automobile notes the mileages that she has obtained in her own urban driving when she fills her automobile's tank with gasoline on nine different occasions. She finds that the results, in miles per gallon, are as follows: 15.6, 18.6, 18.3, 20.1, 21.5, 18.4, 19.1, 20.4, and 19.0. Test the manufacturer's claim by carrying out a test at the level of significance $\alpha_0 = 0.05$. List carefully the assumptions you make.

4. Suppose that a random sample of eight observations $X_1, \ldots, X_8$ is taken from the normal distribution with unknown mean μ and unknown variance σ^2, and it is desired to test the following hypotheses:

$$H_0: \quad \mu = 0,$$
$$H_1: \quad \mu \ne 0.$$

Suppose also that the sample data are such that $\sum_{i=1}^{8} X_i = -11.2$ and $\sum_{i=1}^{8} X_i^2 = 43.7$. If a symmetric t test is performed at the level of significance 0.10 so that each tail of the critical region has probability 0.05, should the hypothesis H_0 be rejected or not?

5. Consider again the conditions of Exercise 4, and suppose again that a t test is to be performed at the level of significance 0.10. Suppose now, however, that the t test is not to be symmetric and the hypothesis H_0 is to be rejected if either $U \le c_1$ or $U \ge c_2$, where $\Pr(U \le c_1) = 0.01$ and $\Pr(U \ge c_2) = 0.09$. For the sample data specified in Exercise 4, should H_0 be rejected or not?

6. Suppose that the variables $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2, and a t test at a given level of significance α_0 is to be carried out to test the following hypotheses:

$$H_0: \quad \mu \le \mu_0,$$
$$H_1: \quad \mu > \mu_0.$$

Let $\pi(\mu, \sigma^2|\delta)$ denote the power function of this t test, and assume that (μ_1, σ_1^2) and (μ_2, σ_2^2) are values of the parameters such that

$$\frac{\mu_1 - \mu_0}{\sigma_1} = \frac{\mu_2 - \mu_0}{\sigma_2}.$$

Show that $\pi(\mu_1, \sigma_1^2|\delta) = \pi(\mu_2, \sigma_2^2|\delta)$.

7. Consider the normal distribution with unknown mean μ and unknown variance σ^2, and suppose that it is desired to test the following hypotheses:

$$H_0: \quad \mu \le \mu_0,$$
$$H_1: \quad \mu > \mu_0.$$

Suppose that it is possible to observe only a single value of X from this distribution, but that an independent random sample of n observations $Y_1, \dots, Y_n$ is available from the normal distribution with known mean 0 and the same variance σ^2 as for X. Show how to carry out a test of the hypotheses H_0 and H_1 based on the t distribution with n degrees of freedom.

8. Suppose that the variables $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Let σ_0^2 be a given positive number, and suppose that it is desired to test the following hypotheses at a specified level of significance α_0 $(0 < \alpha_0 < 1)$:

$$H_0: \quad \sigma^2 \le \sigma_0^2,$$
$$H_1: \quad \sigma^2 > \sigma_0^2.$$

Let $S_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2$, and suppose that the test procedure to be used specifies that H_0 should be rejected if $S_n^2/\sigma_0^2 \ge c$. Also, let $\pi(\mu, \sigma^2|\delta)$ denote the power function of this procedure. Explain how to choose the constant c so that, regardless of the value of μ, the following requirements are satisfied: $\pi(\mu, \sigma^2|\delta) < \alpha_0$ if $\sigma^2 < \sigma_0^2$, $\pi(\mu, \sigma^2|\delta) = \alpha_0$ if $\sigma^2 = \sigma_0^2$, and $\pi(\mu, \sigma^2|\delta) > \alpha_0$ if $\sigma^2 > \sigma_0^2$.

9. Suppose that a random sample of 10 observations $X_1, \dots, X_{10}$ is taken from the normal distribution with unknown mean μ and unknown variance σ^2, and it is desired to test the following hypotheses:

$$H_0: \quad \sigma^2 \le 4,$$
$$H_1: \quad \sigma^2 > 4.$$

Suppose that a test of the form described in Exercise 8 is to be carried out at the level of significance $\alpha_0 = 0.05$. If the observed value of S_n^2 is 60, should the hypothesis H_0 be rejected or not?

10. Suppose again, as in Exercise 9, that a random sample of 10 observations is taken from the normal distribution with unknown mean μ and unknown variance σ^2, but suppose now that the following hypotheses are to be tested at the level of significance 0.05:

$$H_0: \quad \sigma^2 = 4,$$
$$H_1: \quad \sigma^2 \ne 4.$$

Suppose that the null hypothesis H_0 is to be rejected if either $S_n^2 \le c_1$ or $S_n^2 \ge c_2$, where the constants c_1 and c_2 are to be chosen so that, when the hypothesis H_0 is true,

$$\Pr(S_n^2 \le c_1) = \Pr(S_n^2 \ge c_2) = 0.025.$$

Determine the values of c_1 and c_2.

11. Suppose that U_1 has the noncentral t distribution with m degrees of freedom and noncentrality parameter ψ, and suppose that U_2 has the noncentral t distribution with m degrees of freedom and noncentrality parameter $-\psi$. Prove that $\Pr(U_1 \ge c) = \Pr(U_2 \le -c)$.

12. Suppose that a random sample $X_1, \dots, X_n$ is to be taken from the normal distribution with unknown mean μ and unknown variance σ^2, and the following hypotheses are to be tested:

$$H_0: \quad \mu \le 3,$$
$$H_1: \quad \mu > 3.$$

Suppose also that the sample size n is 17, and it is found from the observed values in the sample that $\overline{X}_n = 3.2$ and $(1/n) \sum_{i=1}^n (X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U, and find the corresponding p-value.

13. Consider again the conditions of Exercise 12, but suppose now that the sample size n is 170, and it is again found from the observed values in the sample that $\overline{X}_n = 3.2$ and $(1/n) \sum_{i=1}^n (X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U and find the corresponding p-value.

14. Consider again the conditions of Exercise 12, but suppose now that the following hypotheses are to be tested:

$$H_0: \quad \mu = 3.1,$$
$$H_1: \quad \mu \ne 3.1.$$

Suppose, as in Exercise 12, that the sample size n is 17, and it is found from the observed values in the sample that

$\overline{X}_n = 3.2$ and $(1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U and find the corresponding p-value.

15. Consider again the conditions of Exercise 14, but suppose now that the sample size n is 170, and it is again found from the observed values in the sample that $\overline{X}_n = 3.2$ and $(1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U and find the corresponding p-value.

16. Consider again the conditions of Exercise 14. Suppose, as in Exercise 14, that the sample size n is 17, but suppose now that it is found from the observed values in the sample that $\overline{X}_n = 3.0$ and $(1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U and find corresponding p-value.

17. Prove that the likelihood ratio test for hypotheses (9.5.7) is the two-sided t test that rejects H_0 if $|U| \geq c$, where U is defined in Eq. (8.5.1). The argument is slightly simpler than, but very similar to, the one given in the text for the one-sided case.

18. Prove that the likelihood ratio test for hypotheses (9.5.3) is to reject H_0 if $U \leq c$, where U is defined in Eq. (8.5.1).

9.6 Comparing the Means of Two Normal Distributions

It is very common to compare two distributions to see which has the higher mean or just to see how different the two means are. When the two distributions are normal, the tests and confidence intervals based on the t distribution are very similar to the ones that arose when we considered a single distribution.

The Two-Sample t Test

Example 9.6.1 Rain from Seeded Clouds. In Example 8.3.1, we were interested in whether or not the mean log-rainfall from seeded clouds was greater than 4, which we supposed to have been the mean log-rainfall from unseeded clouds. If we want to compare rainfalls from seeded and unseeded clouds under otherwise similar conditions, we would normally observe two random samples of rainfalls: one from seeded clouds and one from unseeded clouds but otherwise under similar conditions. We would then model these samples as being random samples from two different normal distributions, and we would want to compare their means and possibly their variances to see how different the distributions are. ◄

Consider first a problem in which random samples are available from two normal distributions with common unknown variance, and it is desired to determine which distribution has the larger mean. Specifically, we shall assume that $\boldsymbol{X} = (X_1, \ldots, X_m)$ form a random sample of m observations from a normal distribution for which both the mean μ_1 and the variance σ^2 are unknown, and that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ form an independent random sample of n observations from another normal distribution for which both the mean μ_2 and the variance σ^2 are unknown. We will then be interested in testing hypotheses such as

$$H_0\colon \mu_1 \leq \mu_2 \quad \text{versus} \quad H_1\colon \mu_1 > \mu_2. \tag{9.6.1}$$

For each test procedure δ, we shall let $\pi(\mu_1, \mu_2, \sigma^2|\delta)$ denote the power function of δ. We shall assume that the variance σ^2 is the same for both distributions, even though the value of σ^2 is unknown. If this assumption seems unwarranted, the two-sample t test that we shall derive next would not be appropriate. A different test procedure is discussed later in this section for the case in which the two populations might have different variances. Later in this section, we shall derive the likelihood ratio test. In Sec. 9.7, we discuss some procedures for comparing the variances of two normal distributions, which includes testing the null hypothesis that the variances are the same.

Intuitively, it makes sense to reject H_0 in (9.6.1) if the difference between the sample means is large. Theorem 9.6.1 derives the distribution of a natural test statistic to use.

Theorem 9.6.1 Two-Sample t Statistic. Assume the structure described in the preceding paragraphs. Define

$$\overline{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$S_X^2 = \sum_{i=1}^{m}(X_i - \overline{X}_m)^2, \quad \text{and} \quad S_Y^2 = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2. \tag{9.6.2}$$

Define the test statistic

$$U = \frac{(m+n-2)^{1/2}(\overline{X}_m - \overline{Y}_n)}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}(S_X^2 + S_Y^2)^{1/2}}. \tag{9.6.3}$$

For all values of $\theta = (\mu_1, \mu_2, \sigma^2)$ such that $\mu_1 = \mu_2$, the distribution of U is the t distribution with $m+n-2$ degrees of freedom.

Proof Assume that $\mu_1 = \mu_2$. Define the following two random variables:

$$Z = \frac{\overline{X}_m - \overline{Y}_n}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}\sigma}, \tag{9.6.4}$$

$$W = \frac{S_X^2 + S_Y^2}{\sigma^2}. \tag{9.6.5}$$

The statistic U can now be represented in the form

$$U = \frac{Z}{[W/(m+n-2)]^{1/2}}. \tag{9.6.6}$$

The remainder of the proof consists of proving that Z has the standard normal distribution, that W has the χ^2 distribution with $m+n-2$ degrees of freedom, and that Z and W are independent. The result then follows from Definition 8.4.1, the definition of the family of t distributions.

We have assumed that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent given θ. It follows that every function of $\boldsymbol{X}$ is independent of every function of $\boldsymbol{Y}$. In particular, $(\overline{X}_m, S_X^2)$ is independent of $(\overline{Y}_n, S_Y^2)$. By Theorem 8.3.1, $\overline{X}_m$ and S_X^2 are independent, and $\overline{Y}_n$ and S_Y^2 are also independent. It follows that all four of $\overline{X}_m$, $\overline{Y}_n$, S_X^2, and S_Y^2 are mutually independent. Hence, Z and W are also independent. It also follows from Theorem 8.3.1 that S_X^2/σ^2 and S_Y^2/σ^2 have, respectively, the χ^2 distributions with $m-1$ and $n-1$ degrees of freedom. Hence, W is the sum of two independent random variables with χ^2 distributions and so has the χ^2 distribution with the sum of the two degrees of freedom, namely, $m+n-2$. $\overline{X}_m - \overline{Y}_n$ has the normal distribution with mean $\mu_1 - \mu_2 = 0$ and variance $\sigma^2/n + \sigma^2/m$. It follows that Z has the standard normal distribution. ■

A two-sample t test with level of significance α_0 is the procedure δ that rejects H_0 if $U \geq T^{-1}_{m+n-2}(1-\alpha_0)$. Theorem 9.6.2 states some useful properties of two-sample t tests analogous to those of Theorem 9.5.1. The proof is so similar to that of Theorem 9.5.1 that we shall not present it here.

Theorem 9.6.2 Level and Unbiasedness of Two-Sample t Tests. Let δ be the two-sample t test defined above. The power function $\pi(\mu_1, \mu_2, \sigma^2|\delta)$ has the following properties:

i. $\pi(\mu_1, \mu_2, \sigma^2|\delta) = \alpha_0$ when $\mu_1 = \mu_2$,
ii. $\pi(\mu_1, \mu_2, \sigma^2|\delta) < \alpha_0$ when $\mu_1 < \mu_2$,
iii. $\pi(\mu_1, \mu_2, \sigma^2|\delta) > \alpha_0$ when $\mu_1 > \mu_2$,
iv. $\pi(\mu_1, \mu_2, \sigma^2|\delta) \to 0$ as $\mu_1 - \mu_2 \to -\infty$,
v. $\pi(\mu_1, \mu_2, \sigma^2|\delta) \to 1$ as $\mu_1 - \mu_2 \to \infty$.

Furthermore, the test δ has size α_0 and is unbiased. ■

Note: The Other One-Sided Hypotheses. If the hypotheses are

$$H_0\colon \mu_1 \geq \mu_2 \quad \text{versus} \quad H_1\colon \mu_1 < \mu_2, \tag{9.6.7}$$

the corresponding level α_0 t test is to reject H_0 when $U \leq -T^{-1}_{m+n-2}(1-\alpha_0)$. This test has properties analogous to those of the other one-sided test.

P-values are computed in much the same way as they were for the one-sample t test. The proof of Theorem 9.6.3 is virtually the same as the proof of Theorem 9.5.2 and is not given here.

Theorem 9.6.3 p-Values for Two-Sample t Tests. Suppose that we are testing either the hypotheses in Eq. (9.6.1) or the hypotheses in Eq. (9.6.7). Let u be the observed value of the statistic U in Eq. (9.6.3), and let $T_{m+n-2}(\cdot)$ be the c.d.f. of the t distribution with $m+n-2$ degrees of freedom. Then the p-value for the hypotheses in Eq. (9.6.1) is $1 - T_{m+n-2}(u)$ and the p-value for the hypotheses in Eq. (9.6.7) is $T_{m+n-2}(u)$. ■

Example 9.6.2 Rain from Seeded Clouds. In Example 9.6.1, we actually have 26 observations of unseeded clouds to go with the 26 observations of seeded clouds. Let $X_1, \dots, X_{26}$ be the log-rainfall measurements from the seeded clouds, and let $Y_1, \dots, Y_{26}$ be the measurements from the unseeded clouds. We model all of the measurements as independent with the X_i's having the normal distribution with mean μ_1 and variance σ^2, and the Y_i's having the normal distribution with mean μ_2 and variance σ^2. For now, we model the two distributions as having a common variance. Suppose that we wish to test whether or not the mean log-rainfall from seeded clouds is larger than the mean log-rainfall from unseeded clouds. We choose the null and alternative hypotheses so that type I error corresponds to claiming that seeding increases rainfall when, in fact, it does not increase rainfall. That is, the null hypothesis is $H_0 : \mu_1 \leq \mu_2$ and the alternative hypothesis is $H_1 : \mu_1 > \mu_2$. We choose a level of significance of $\alpha_0 = 0.01$. Before proceeding with the formal test, it is a good idea to look at the data first. Figure 9.15 contains histograms of the log-rainfalls of both seeded and unseeded clouds. The two samples look different, with the seeded clouds appearing to have larger log-rainfalls. The formal test requires us to compute the statistics

$$\overline{X}_m = 5.13, \qquad \overline{Y}_n = 3.99,$$
$$S_X^2 = 63.96, \quad \text{and} \quad S_Y^2 = 67.39.$$

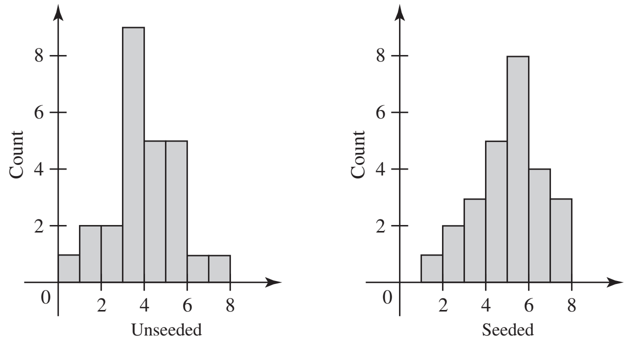


Figure 9.15 Histograms of seeded and unseeded clouds in Example 9.6.2.

The critical value is $T_{50}^{-1}(0.99) = 2.403$, and the test statistic is

$$U = \frac{50^{1/2}(5.13 - 3.99)}{\left(\frac{1}{26} + \frac{1}{26}\right)^{1/2}(63.96 + 67.39)^{1/2}} = 2.544,$$

which is greater than 2.403. So, we would reject the null hypothesis at level of significance $\alpha_0 = 0.01$. The p-value is the smallest level at which we would reject H_0, namely, $1 - T_{50}(2.544) = 0.007$. ◀

Example 9.6.3 Roman Pottery in Britain. Tubb, Parker, and Nickless (1980) describe a study of samples of pottery from the Roman era found in various locations in Great Britain. One measurement made on each sample of pottery was the percentage of the sample that was aluminum oxide. Suppose that we are interested in comparing the aluminum oxide percentages at two different locations. There were $m = 14$ samples analyzed from Llanederyn, with sample average of $\overline{X}_m = 12.56$ and $S_X^2 = 24.65$. Another $n = 5$ samples came from Ashley Rails, with $\overline{Y}_n = 17.32$ and $S_Y^2 = 11.01$. One of the sample sizes is too small for the histogram to be very illuminating. Suppose that we model the data as normal random variables with two different means μ_1 and μ_2 but common variance σ^2. We want to test the null hypothesis $H_0 : \mu_1 \geq \mu_2$ against the alternative hypothesis $H_1 : \mu_1 < \mu_2$. The observed value of U defined by Eq. (9.6.3) is -6.302. From the table of the t distribution in this book, with $m + n - 2 = 17$ degrees of freedom, we find that $T_{17}^{-1}(0.995) = 2.898$ and $U < -2.898$. So, we would reject H_0 at any level $\alpha_0 \geq 0.005$. Indeed, the p-value associated with this value of U is $T_{17}(-6.302) = 4 \times 10^{-6}$. ◀

Power of the Test

For each parameter vector $\theta = (\mu_1, \mu_2, \sigma^2)$, the power function of the two-sample t test can be computed using the noncentral t distribution introduced in Definition 9.5.1. Almost identical reasoning to that which led to Theorem 9.5.3 proves the following.

Theorem 9.6.4 Power of Two-Sample t Test. Assume the conditions stated earlier in this section. Let U be defined in Eq. (9.6.6). Then U has the noncentral t distribution with $m + n - 2$ degrees of freedom and noncentrality parameter

$$\psi = \frac{\mu_1 - \mu_2}{\sigma\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}}. \tag{9.6.8}$$

We can use Fig. 9.12 on page 580 to approximate power calculations if we do not have an appropriate computer program handy.

Example 9.6.4 Roman Pottery in Britain. In Example 9.6.3, if the Llanederyn mean is less than the Ashley Rails mean by 1.5σ, then $\psi = 1.5/(1/14 + 1/5)^{1/2} = 2.88$. The power of a level 0.01 test of $H_0 : \mu_1 \geq \mu_2$ appears to be about 0.65 in the right panel of Fig. 9.12. (The actual power is 0.63.) ◄

Two-Sided Alternatives

The two-sample t test can easily be adapted to testing the following hypotheses at a specified level of significance α_0:

$$H_0\colon \mu_1 = \mu_2, \quad \text{versus} \quad H_1\colon \mu_1 \neq \mu_2. \tag{9.6.9}$$

The size α_0 two-sided t test rejects H_0 if $|U| \geq c$ where $c = T^{-1}_{m+n-2}(1 - \alpha_0/2)$, and the statistic U is defined in Eq. (9.6.3). The p-value when $U = u$ is observed equals $2[1 - T_{m+n-2}(|u|)]$. (See Exercise 9.)

Example 9.6.5 Comparing Copper Ores. Suppose that a random sample of eight specimens of ore is collected from a certain location in a copper mine, and the amount of copper in each of the specimens is measured in grams. We shall denote these eight amounts by $X_1, \ldots, X_8$ and shall suppose that the observed values are such that $\overline{X}_8 = 2.6$ and $S_X^2 = 0.32$. Suppose also that a second random sample of 10 specimens of ore is collected from another part of the mine. We shall denote the amounts of copper in these specimens by $Y_1, \ldots, Y_{10}$ and shall suppose that the observed values in grams are such that $\overline{Y}_{10} = 2.3$, and $S_Y^2 = 0.22$. Let μ_1 denote the mean amount of copper in all the ore at the first location in the mine, let μ_2 denote the mean amount of copper in all the ore at the second location, and suppose that the hypotheses (9.6.9) are to be tested.

We shall assume that all the observations have a normal distribution, and the variance is the same at both locations in the mine, even though the means may be different. In this example, the sample sizes are $m = 8$ and $n = 10$, and the value of the statistic U defined by Eq. (9.6.3) is 3.442. Also, by the use of a table of the t distribution with 16 degrees of freedom, it is found that $T^{-1}_{16}(0.995) = 2.921$, so that the tail area corresponding to this observed value of U is less than 2×0.005. Hence, the null hypothesis will be rejected for any specified level of significance $\alpha_0 \geq 0.01$. (In fact, the two-sided tail area associated with $U = 3.442$ is 0.003.) ◄

The power function of the two-sided two-sample t test is based on the noncentral t distribution in the same way as was the power function of the one-sample two-sided t test. The test δ that rejects $H_0 : \mu_1 = \mu_2$ when $|U| \geq c$ has power function

$$\pi(\mu_1, \mu_2, \sigma^2|\delta) = T_{m+n-2}(-c|\psi) + 1 - T_{m+n-2}(c|\psi),$$

where $T_{m+n-2}(\cdot|\psi)$ is the c.d.f. of the noncentral t distribution with $m + n - 2$ degrees of freedom and noncentrality parameter ψ given in Eq. (9.6.8). Figure 9.14 on page 583 can be used to approximate the power function if appropriate software is not available.

The Two-Sample t Test as a Likelihood Ratio Test

In this section, we shall show that the two-sample t test for the hypotheses (9.6.1) is a likelihood ratio test. After the values $x_1, \dots, x_m$ and $y_1, \dots, y_n$ in the two samples have been observed, the likelihood function $g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2)$ is

$$g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2) = f_m(\boldsymbol{x}|\mu_1, \sigma^2) f_n(\boldsymbol{y}|\mu_2, \sigma^2).$$

Here, both $f_m(\boldsymbol{x}|\mu_1, \sigma^2)$ and $f_n(\boldsymbol{y}|\mu_2, \sigma^2)$ have the form given in Eq. (9.5.9), and the value of σ^2 is the same in both terms. In this case, $\Omega_0 = \{(\mu_1, \mu_2, \sigma^2) : \mu_1 \le \mu_2\}$. The likelihood ratio statistic is

$$\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sup_{\{(\mu_1,\mu_2,\sigma^2):\mu_1 \le \mu_2\}} g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2)}{\sup_{(\mu_1,\mu_2,\sigma^2)} g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2)}. \tag{9.6.10}$$

The likelihood ratio test procedure then specifies that H_0 should be rejected if $\Lambda(\boldsymbol{x}, \boldsymbol{y}) \le k$, where k is typically chosen so that the test has a desired level α_0.

To facilitate the maximizations in (9.6.10), let

$$s_x^2 = \sum_{i=1}^{m} (x_i - \overline{x}_m)^2, \text{ and } s_y^2 = \sum_{i=1}^{n} (y_i - \overline{y}_n)^2.$$

Then we can write

$$\begin{aligned} &g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2) \\ &\quad = \frac{1}{(2\pi\sigma^2)^{(m+n)/2}} \exp\left(-\frac{1}{2\sigma^2}\left[m(\overline{x}_m - \mu_1)^2 + n(\overline{y}_n - \mu_2)^2 + s_x^2 + s_y^2\right]\right). \end{aligned}$$

The denominator of (9.6.10) is maximized by the overall M.L.E.'s, that is, when

$$\mu_1 = \overline{x}_m, \qquad \mu_2 = \overline{y}_n, \quad \text{and} \quad \sigma^2 = \frac{1}{m+n}(s_x^2 + s_y^2). \tag{9.6.11}$$

For the numerator of (9.6.10), when $\overline{x}_m \le \overline{y}_n$, the parameter vector in (9.6.11) is in Ω_0, and hence the maximum also occurs at the values in Eq. (9.6.11). Hence, $\Lambda(\boldsymbol{x}, \boldsymbol{y}) = 1$ if $\overline{x}_m \le \overline{y}_n$.

For the other case, when $\overline{x}_m > \overline{y}_n$, it is not difficult to see that $\mu_1 = \mu_2$ is required in order to achieve the maximum. In these cases, the maximum occurs when

$$\begin{aligned} \mu_1 = \mu_2 &= \frac{m\overline{x}_m + n\overline{y}_n}{m+n}, \\ \sigma^2 &= \frac{mn(\overline{x}_m - \overline{y}_n)^2/(m+n) + s_x^2 + s_y^2}{m+n}. \end{aligned}$$

Substituting all of these values into (9.6.10) yields

$$\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 1 & \text{if } \overline{x}_m \le \overline{y}_n, \\ (1+v^2)^{-(m+n)/2} & \text{if } \overline{x}_m > \overline{y}_n, \end{cases}$$

where

$$v = \frac{(\overline{x}_m - \overline{y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2} (s_x^2 + s_y^2)^{1/2}}. \tag{9.6.12}$$

If $k < 1$, it is straightforward to show that $\Lambda(\boldsymbol{x}, \boldsymbol{y}) \le k$ is equivalent to $v \ge k'$ for some other constant k'. Finally, note that $(m+n-2)^{1/2} v$ is the observed value of U, so the likelihood ratio test is to reject H_0 when $U \geq c$, for some constant c. This is the same as the two-sample t test. The preceding argument can easily be adapted to handle the other one-sided hypotheses and the two-sided case. (See Exercise 13 for the two-sided case.)

Unequal Variances

Known Ratio of Variances The t test can be extended to a problem in which the variances of the two normal distributions are not equal but the ratio of one variance to the other is known. Specifically, suppose that $X_1, \dots, X_m$ form a random sample from the normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, \dots, Y_n$ form an independent random sample from another normal distribution with mean μ_2 and variance σ_2^2. Suppose also that the values of μ_1, μ_2, σ_1^2, and σ_2^2 are unknown but that $\sigma_2^2 = k\sigma_1^2$, where k is a known positive constant. Then it can be shown (see Exercise 4 at the end of this section) that when $\mu_1 = \mu_2$, the following random variable U will have the t distribution with $m + n - 2$ degrees of freedom:

$$U = \frac{(m+n-2)^{1/2}(\overline{X}_m - \overline{Y}_n)}{\left(\dfrac{1}{m} + \dfrac{k}{n}\right)^{1/2}\left(S_X^2 + \dfrac{S_Y^2}{k}\right)^{1/2}}. \tag{9.6.13}$$

Hence, the statistic U defined by Eq. (9.6.13) can be used for testing either the hypotheses (9.6.1) or the hypotheses (9.6.9).

The Behrens-Fisher Problem If the values of all four parameters μ_1, μ_2, σ_1^2, and σ_2^2 are unknown, and if the value of the ratio σ_1^2/σ_2^2 is also unknown, then the problem of testing the hypotheses (9.6.1) or the hypotheses (9.6.9) becomes very difficult. Even the likelihood ratio statistic Λ has no known distribution. This problem is known as the *Behrens-Fisher problem*. Some simulation methods for the Behrens-Fisher problem will be described in Chapter 12 (Examples 12.2.4 and 12.6.10). Various other test procedures have been proposed, but most of them have been the subject of controversy in regard to their appropriateness or usefulness. The most popular of the proposed methods was developed in a series of articles by Welch (1938, 1947, 1951). Welch proposed using the statistic

$$V = \frac{\overline{X}_m - \overline{Y}_n}{\left(\dfrac{S_X^2}{m(m-1)} + \dfrac{S_Y^2}{n(n-1)}\right)^{1/2}}. \tag{9.6.14}$$

Even when $\mu_1 = \mu_2$, the distribution of V is not known in closed form. However, Welch approximated the distribution of V by a t distribution as follows. Let

$$W = \frac{S_X^2}{m(m-1)} + \frac{S_Y^2}{n(n-1)}, \tag{9.6.15}$$

and approximate the distribution of W by a gamma distribution with the same mean and variance as W. (See Exercise 12.) If we were now to assume that W actually had this approximating gamma distribution, then V would have the t distribution with degrees of freedom

$$\frac{\left(\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{\sigma_1^2}{m}\right)^2+\frac{1}{n-1}\left(\frac{\sigma_2^2}{n}\right)^2}. \tag{9.6.16}$$

Next, substitute the unbiased estimates $s_x^2/(m-1)$ and $s_y^2/(n-1)$ for σ_1^2 and σ_2^2, respectively, in (9.6.16) to obtain the degrees of freedom for Welch's t distribution approximation:

$$\nu=\frac{\left(\frac{s_x^2}{m(m-1)}+\frac{s_y^2}{n(n-1)}\right)^2}{\frac{1}{(m-1)^3}\left(\frac{s_x^2}{m}\right)^2+\frac{1}{(n-1)^3}\left(\frac{s_y^2}{n}\right)^2}. \tag{9.6.17}$$

In Eq. (9.6.17), s_x^2 and s_y^2 are the observed values of S_X^2 and S_Y^2. To summarize Welch's procedure, act as if V in Eq. (9.6.14) had the t distribution with ν degrees of freedom when $\mu_1=\mu_2$. Tests of one-sided and two-sided hypotheses are then constructed by comparing V to various quantiles of the t distribution with ν degrees of freedom. If ν is not an integer, round it to the nearest integer or use a computer program that can handle t distributions with noninteger degrees of freedom.

Example 9.6.6

Comparing Copper Ores. Using the data from Example 9.6.5, we compute

$$V=\frac{2.6-2.3}{\left(\frac{0.32}{8\times 7}+\frac{0.22}{10\times 9}\right)^{1/2}}=3.321,$$

$$\nu=\frac{\left(\frac{0.32}{8\times 7}+\frac{0.22}{10\times 9}\right)^2}{\frac{1}{7^3}\left(\frac{0.32}{8}\right)^2+\frac{1}{9^3}\left(\frac{0.22}{10}\right)^2}=12.49.$$

The p-value associated with the observed data for the hypotheses (9.6.9) is $2[1-T_{12.49}(3.321)]=0.0058$, not much different than what we obtained in Example 9.6.5. ◄

Likelihood Ratio Test An alternative to the Welch approximation described above would be to apply the large-sample approximation of Theorem 9.1.4. Using the same notation as earlier in the section, we can write the likelihood function as

$$\begin{aligned} &g(\boldsymbol{x},\boldsymbol{y}|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) \\ &\quad=\frac{1}{(2\pi\sigma_1^2)^{m/2}(2\pi\sigma_2^2)^{n/2}}\exp\left(-\frac{m(\overline{x}_m-\mu_1)^2+s_x^2}{2\sigma_1^2}-\frac{n(\overline{y}_n-\mu_2)^2+s_y^2}{2\sigma_2^2}\right). \end{aligned} \tag{9.6.18}$$

The overall M.L.E.'s are

$$\hat{\mu}_1=\overline{x}_m,\ \ \hat{\mu}_2=\overline{y}_n,\ \ \hat{\sigma}_1^2=\frac{s_x^2}{m},\ \hat{\sigma}_2^2=\frac{s_y^2}{n}. \tag{9.6.19}$$

Under $H_0 : \mu_1 = \mu_2$, we cannot find formulas for the M.L.E.'s. However, if we let $\hat{\mu}$ stand for the common value of $\hat{\mu}_1 = \hat{\mu}_2$, we find that the M.L.E.'s satisfy the following equations:

$$\hat{\sigma}_1^2 = \frac{1}{m}\left[s_x^2 + m(\overline{x}_m - \hat{\mu})^2\right], \tag{9.6.20}$$

$$\hat{\sigma}_2^2 = \frac{1}{n}\left[s_y^2 + n(\overline{y}_n - \hat{\mu})^2\right], \tag{9.6.21}$$

$$\hat{\mu} = \frac{\dfrac{m\overline{x}_m}{\hat{\sigma}_1^2} + \dfrac{n\overline{y}_n}{\hat{\sigma}_2^2}}{\dfrac{m}{\hat{\sigma}_1^2} + \dfrac{n}{\hat{\sigma}_2^2}}. \tag{9.6.22}$$

These equations can be solved recursively even though we do not have a closed-form solution. One algorithm is the following:

1. Set $k = 0$ and pick a starting value $\hat{\mu}^{(0)}$, such as $(m\overline{x}_m + n\overline{y}_n)/(m + n)$.
2. Compute $\hat{\sigma}_1^{2(k)}$ and $\hat{\sigma}_2^{2(k)}$ by substituting $\hat{\mu}^{(k)}$ into Eqs. (9.6.20) and (9.6.21).
3. Compute $\hat{\mu}^{(k+1)}$ by substituting $\hat{\sigma}_1^{2(k)}$ and $\hat{\sigma}_2^{2(k)}$ into Eq. (9.6.22).
4. If $\hat{\mu}^{(k+1)}$ is close enough to $\hat{\mu}^{(k)}$ stop. Otherwise, replace k by $k + 1$ and return to step 2.

Example 9.6.7 Comparing Copper Ores. Using the data in Example 9.6.5, we will start with $\hat{\mu}^{(0)} = (8 \times 2.6 + 10 \times 2.3)/18 = 2.433$. Plugging this value into Eqs. (9.6.20) and (9.6.21) gives us $\hat{\sigma}_1^{2(0)} = 0.068$ and $\hat{\sigma}_2^{2(0)} = 0.0398$. Plugging these into Eq. (9.6.22) gives $\hat{\mu}^{(1)} = 2.396$. After 13 iterations the values stop changing and our final M.L.E.'s are $\hat{\mu} = 2.347$, $\hat{\sigma}_1^2 = 0.1039$, and $\hat{\sigma}_2^2 = 0.0242$. We can then substitute these M.L.E.'s into the likelihood function (9.6.18) to get the numerator of the likelihood ratio statistic $\Lambda(\boldsymbol{x}, \boldsymbol{y})$. (Remember to substitute $\hat{\mu}$ for both μ_1 and μ_2.) We can also substitute the overall M.L.E.'s (9.6.19) into (9.6.18) to get the denominator of $\Lambda(\boldsymbol{x}, \boldsymbol{y})$. The result is $\Lambda(\boldsymbol{x}, \boldsymbol{y}) = 0.01356$. Theorem 9.1.4 says that we should compare $-2 \log \Lambda(\boldsymbol{x}, \boldsymbol{y}) = 8.602$ to a critical value of the χ^2 distribution with one degree of freedom. The p-value associated with the observed statistic is the probability that a χ^2 random variable with one degree of freedom is greater than 8.602, namely, 0.003. This is the same as the p-value that we obtained in Example 9.6.5 when we assumed that the two variances were the same. ◄

For the cases of one-sided hypotheses such as (9.6.1) and (9.6.7), the likelihood ratio statistic is a bit more complicated. For example, if $\mu_1 = \mu_2$, $-2 \log \Lambda(\boldsymbol{X}, \boldsymbol{Y})$ converges in distribution to a distribution that is neither discrete nor continuous. We will not discuss this case further in this book.

Summary

Suppose that we observe independent random samples from two normal distributions: $X_1, \dots, X_m$ having mean μ_1 and variance σ_1^2, and $Y_1, \dots, Y_n$ having mean μ_2 and variance σ_2^2. For testing hypotheses about μ_1 and μ_2, t tests are available if we assume that $\sigma_1^2 = \sigma_2^2$. The t tests all make use of the statistic U defined in Eq. (9.6.3). To test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ at level α_0, reject H_0 if $|U| \geq$

$T^{-1}_{m+n-2}(1-\alpha_0/2)$, where T^{-1}_{m+n-2} is the quantile function of the t distribution with $m+n-2$ degrees of freedom. To test $H_0: \mu_1 \le \mu_2$ versus $H_1: \mu_1 > \mu_2$ at level α_0, reject H_0 if $U > T^{-1}_{m+n-2}(1-\alpha_0)$. To test $H_0: \mu_1 \ge \mu_2$ versus $H_1: \mu_1 < \mu_2$ at level α_0, reject H_0 if $U < -T^{-1}_{m+n-2}(1-\alpha_0)$. The power functions of these tests can be computed using the family of noncentral t distributions. Approximate tests are available if we do not assume that $\sigma_1^2 = \sigma_2^2$.

Exercises

1. In Example 9.6.3, we discussed Roman pottery found at two different locations in Great Britain. There were samples found at other locations as well. One other location, Island Thorns, had five samples $X_1, \dots, X_n$ with an average aluminum oxide percentage of $\overline{X} = 18.18$ with $\sum_{i=1}^{5}(X_i - \overline{X})^2 = 12.61$. Let $Y_1, \dots, Y_5$ be the five sample measurements from Ashley Rails in Example 9.6.3. Test the null hypothesis that the mean aluminum oxide percentages at Ashely Rails and Island Thorns are the same versus the alternative that they are different at level $\alpha_0 = 0.05$.

2. Suppose that a certain drug A was administered to eight patients selected at random, and after a fixed time period, the concentration of the drug in certain body cells of each patient was measured in appropriate units. Suppose that these concentrations for the eight patients were found to be as follows:

1.23, 1.42, 1.41, 1.62, 1.55, 1.51, 1.60, and 1.76.

Suppose also that a second drug B was administered to six different patients selected at random, and when the concentration of drug B was measured in a similar way for these six patients, the results were as follows:

1.76, 1.41, 1.87, 1.49, 1.67, and 1.81.

Assuming that all the observations have a normal distribution with a common unknown variance, test the following hypotheses at the level of significance 0.10: The null hypothesis is that the mean concentration of drug A among all patients is at least as large as the mean concentration of drug B. The alternative hypothesis is that the mean concentration of drug B is larger than that of drug A.

3. Consider again the conditions of Exercise 2, but suppose now that it is desired to test the following hypotheses: The null hypothesis is that the mean concentration of drug A among all patients is the same as the mean concentration of drug B. The alternative hypothesis, which is two-sided, is that the mean concentrations of the two drugs are not the same. Find the number c so that the level 0.05 two-sided t test will reject H_0 when $|U| \ge c$, where U is defined by Eq. (9.6.3). Also, perform the test.

4. Suppose that $X_1, \dots, X_m$ form a random sample from the normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, \dots, Y_n$ form an independent random sample from the normal distribution with mean μ_2 and variance σ_2^2. Show that if $\mu_1 = \mu_2$ and $\sigma_2^2 = k\sigma_1^2$, then the random variable U defined by Eq. (9.6.13) has the t distribution with $m+n-2$ degrees of freedom.

5. Consider again the conditions and observed values of Exercise 2. However, suppose now that each observation for drug A has an unknown variance σ_1^2, and each observation for drug B has an unknown variance σ_2^2, but it is known that $\sigma_2^2 = (6/5)\sigma_1^2$. Test the hypotheses described in Exercise 2 at the level of significance 0.10.

6. Suppose that $X_1, \dots, X_m$ form a random sample from the normal distribution with unknown mean μ_1 and unknown variance σ^2, and $Y_1, \dots, Y_n$ form an independent random sample from another normal distribution with unknown mean μ_2 and the same unknown variance σ^2. For each constant λ ($-\infty < \lambda < \infty$), construct a t test of the following hypotheses with $m+n-2$ degrees of freedom:

$$\begin{aligned} H_0&: \quad \mu_1 - \mu_2 = \lambda, \\ H_1&: \quad \mu_1 - \mu_2 \ne \lambda. \end{aligned}$$

7. Consider again the conditions of Exercise 2. Let μ_1 denote the mean of each observation for drug A, and let μ_2 denote the mean of each observation for drug B. It is assumed, as in Exercise 2, that all the observations have a common unknown variance. Use the results of Exercise 6 to construct a confidence interval for $\mu_1 - \mu_2$ with confidence coefficient 0.90.

8. In Example 9.6.5, determine the power of a level 0.01 test if $|\mu_1 - \mu_2| = \sigma$.

9. Suppose that we wish to test the hypotheses (9.6.9). We shall use the statistic U defined in Eq. (9.6.3) and reject H_0 if $|U|$ is large. Prove that the p-value when $U = u$ is observed is $2[1 - T_{m+n-2}(|u|)]$.

10. Lyle et al. (1987) ran an experiment to study the effect of a calcium supplement on the blood pressure of African American males. A group of 10 men received a calcium supplement, and another group of 11 men received a placebo. The experiment lasted 12 weeks. Both

Table 9.2 Blood pressure data for Exercise 10

Calcium	7	−4	18	17	−3	−5	1	10	11	−2	
Placebo	−1	12	−1	−3	3	−5	5	2	−11	−1	−3

before and after the 12-week period, each man had his systolic blood pressure measured while at rest. The changes (after minus before) are given in Table 9.2. Test the null hypothesis that the mean change in blood pressure for the calcium supplement group is lower than the mean change in blood pressure for the placebo group. Use level $\alpha_0 = 0.1$.

11. Frisby and Clatworthy (1975) studied the times that it takes subjects to fuse random-dot stereograms. Random-dot stereograms are pairs of images that appear at first to be random dots. After a subject looks at the pair of images from the proper distance and her eyes cross just the right amount, a recognizable object appears from the fusion of the two images. The experimenters were concerned with the extent to which prior information about the recognizable object affected the time it took to fuse the images.

One group of 43 subjects was not shown a picture of the object before being asked to fuse the images. Their average time was $\overline{X}_{43} = 8.560$ and $S_X^2 = 2745.7$. The second group of 35 subjects was shown a picture of the object, and their sample statistics were $\overline{Y}_{35} = 5.551$ and $S_Y^2 = 783.9$. The null hypothesis is that the mean time of the first group is no larger than the mean time of the second group, while the alternative hypothesis is that the first group takes longer.

a. Test the hypotheses at the level of significance $\alpha_0 = 0.01$, assuming that the variances are equal for the two groups.

b. Test the hypotheses at the level of significance $\alpha_0 = 0.01$, using Welch's approximate test.

12. Find the mean a and variance b of the random variable W in Eq. (9.6.15). Now, let a and b be the mean and variance, respectively, of the gamma distribution with parameters α and β. Prove that 2α equals the expression in (9.6.16).

13. Let U be as defined in Eq. (9.6.3), and suppose that it is desired to test the hypotheses in Eq. (9.6.9). Prove that each likelihood ratio test has the following form: reject H_0 if $|U| \geq c$, where c is a constant. *Hint:* First prove that $\Lambda(\boldsymbol{x}, \boldsymbol{y}) = (1 + v^2)^{-(m+n)/2}$, where v was defined in Eq. (9.6.12).

9.7 The F Distributions

In this section, we introduce the family of F distributions. This family is useful in two different hypothesis-testing situations. The first situation is when we wish to test hypotheses about the variances of two different normal distributions. These tests, which we shall derive in this section, are based on a statistic that has an F distribution. The second situation will arise in Chapter 11 when we test hypotheses concerning the means of more than two normal distributions.

Definition of the F Distribution

Example 9.7.1 Rain from Seeded Clouds. In Example 9.6.1, we were interested in comparing the distributions of log-rainfalls from seeded and unseeded clouds. In Example 9.6.2, we used the two-sample t test to compare the means of these distributions under the assumption that the variances of the two distributions were the same. It would be good to have a procedure for testing whether or not such an assumption is warranted. ◄

In this section, we shall introduce a family of distributions, called the F distributions, that arises in many important problems of testing hypotheses in which two or more normal distributions are to be compared on the basis of random samples from each of the distributions. In particular, it arises naturally when we wish to compare the variances of two normal distributions.

Definition 9.7.1 The F distributions. Let Y and W be independent random variables such that Y has the χ^2 distribution with m degrees of freedom and W has the χ^2 distribution with n degrees of freedom, where m and n are given positive integers. Define a new random variable X as follows:

$$X = \frac{Y/m}{W/n} = \frac{nY}{mW}. \tag{9.7.1}$$

Then the distribution of X is called the *F distribution with m and n degrees of freedom*.

Theorem 9.7.1 gives the general p.d.f. of an F distribution. Its proof relies on the methods of Sec. 3.9 and will be postponed until the end of this section.

Theorem 9.7.1 Probability Density Function. Let X have the F distribution with m and n degrees of freedom. Then its p.d.f. $f(x)$ is as follows, for $x > 0$:

$$f(x) = \frac{\Gamma\left[\frac{1}{2}(m+n)\right] m^{m/2} n^{n/2}}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)} \cdot \frac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}}, \tag{9.7.2}$$

and $f(x) = 0$ for $x \leq 0$.

Properties of the F Distributions

When we speak of the F distribution with m and n degrees of freedom, the order in which the numbers m and n are given is important, as can be seen from the definition of X in Eq. (9.7.1). When $m \neq n$, the F distribution with m and n degrees of freedom and the F distribution with n and m degrees of freedom are two different distributions. Theorem 9.7.2 gives a result relating the two distributions just mentioned along with a relationship between F distributions and t distributions.

Theorem 9.7.2 If X has the F distribution with m and n degrees of freedom, then its reciprocal $1/X$ has the F distribution with n and m degrees of freedom. If Y has the t distribution with n degrees of freedom, then Y^2 has the F distribution with 1 and n degrees of freedom.

Proof The first statement follows from the representation of X as the ratio of two random variables, in Definition 9.7.1. The second statement follows from the representation of a t random variable in the form of Eq. (8.4.1). ∎

Two short tables of quantiles for F distributions are given at the end of this book. In these tables, we give only the 0.95 quantile and the 0.975 quantile for different possible pairs of values of m and n. In other words, if G denotes the c.d.f. of the F distribution with m and n degrees of freedom, then the tables give the values of x_1 and x_2 such that $G(x_1) = 0.95$ and $G(x_2) = 0.975$. By applying Theorem 9.7.2, it is possible to use the tables to obtain the 0.05 and 0.025 quantiles of an F distribution. Most statistical software will compute the c.d.f. and quatiles for general F distributions.

Example 9.7.2 Determining the 0.05 Quantile of an F Distribution. Suppose that a random variable X has the F distribution with 6 and 12 degrees of freedom. We shall determine the 0.05 quantile of X, that is, the value of x such that $\Pr(X < x) = 0.05$.

If we let $Y = 1/X$, then Y will have the F distribution with 12 and 6 degrees of freedom. It can be found from the table given at the end of this book that $\Pr(Y \le 4.00) = 0.95$; hence, $\Pr(Y > 4.00) = 0.05$. Since $Y > 4.00$ if and only if $X < 0.25$, it follows that $\Pr(X < 0.25) = 0.05$. Because F distributions are continuous, $\Pr(X \le 0.25) = 0.05$, and 0.25 is the 0.05 quantile of X. ◄

Comparing the Variances of Two Normal Distributions

Suppose that the random variables $X_1, \dots, X_m$ form a random sample of m observations from a normal distribution for which both the mean μ_1 and the variance σ_1^2 are unknown, and suppose also that the random variables $Y_1, \dots, Y_n$ form an independent random sample of n observations from another normal distribution for which both the mean μ_2 and the variance σ_2^2 are unknown. Suppose finally that the following hypotheses are to be tested at a specified level of significance α_0 $(0 < \alpha_0 < 1)$:

$$\begin{aligned} H_0&: \quad \sigma_1^2 \le \sigma_2^2, \\ H_1&: \quad \sigma_1^2 > \sigma_2^2. \end{aligned} \tag{9.7.3}$$

For each test procedure δ, we shall let $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta)$ denote the power function of δ. Later in this section, we shall derive the likelihood ratio test. For now, define S_X^2 and S_Y^2 to be the sums of squares defined in Eq. (9.6.2). Then $S_X^2/(m-1)$ and $S_Y^2/(n-1)$ are estimators of σ_1^2 and σ_2^2, respectively. It makes intuitive sense that we should reject H_0 if the ratio of these two estimators is large. That is, define

$$V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)}, \tag{9.7.4}$$

and reject H_0 if $V \ge c$, where c is chosen to make the test have a desired level of significance.

Definition 9.7.2 *F* test. The test procedure defined above is called an F *test*.

Properties of F Tests

Theorem 9.7.3 Distribution of V. Let V be the statistic in Eq. (9.7.4). The distribution of $(\sigma_2^2/\sigma_1^2)V$ is the F distribution with $m-1$ and $n-1$ degrees of freedom. In particular, if $\sigma_1^2 = \sigma_2^2$, then the distribution of V itself is the F distribution with $m-1$ and $n-1$ degrees of freedom.

Proof We know from Theorem 8.3.1 that the random variable S_X^2/σ_1^2 has the χ^2 distribution with $m-1$ degrees of freedom, and the random variable S_Y^2/σ_2^2 has the χ^2 distribution with $n-1$ degrees of freedom. Furthermore, these two random variables are independent, since they are calculated from two independent samples. Therefore, the following random variable V^* has the F distribution with $m-1$ and $n-1$ degrees of freedom:

$$V^* = \frac{S_X^2/[(m-1)\sigma_1^2]}{S_Y^2/[(n-1)\sigma_2^2]}. \tag{9.7.5}$$

It can be seen from Eqs. (9.7.4) and (9.7.5) that $V^* = (\sigma_2^2/\sigma_1^2)V$. This proves the first claim in the theorem. If $\sigma_1^2 = \sigma_2^2$, then $V = V^*$, which proves the second claim. ■

If $\sigma_1^2 = \sigma_2^2$, it is possible to use a table of the F distribution to choose a constant c such that $\Pr(V \ge c) = \alpha_0$, regardless of the common value of σ_1^2 and σ_2^2, and regardless of the values of μ_1 and μ_2. In fact, c will be the $1 - \alpha_0$ quantile of the corresponding F distribution. We prove next that the test that rejects H_0 in (9.7.3) if $V \ge c$ has level α_0.

Theorem 9.7.4 Level, Power Function, and P-Values. Let V be the statistic defined in Eq. (9.7.4). Let c be the $1 - \alpha_0$ quantile of the F distribution with $m - 1$ and $n - 1$ degrees of freedom, and let $G_{m-1,n-1}$ be the c.d.f. of that F distribution. Let δ be test that rejects H_0 in (9.7.3) when $V \ge c$. The power function $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta)$ satisfies the following properties:

i. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) = 1 - G_{m-1,n-1}\left(\frac{\sigma_2^2}{\sigma_1^2}c\right)$,

ii. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) = \alpha_0$ when $\sigma_1^2 = \sigma_2^2$,

iii. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) < \alpha_0$ when $\sigma_1^2 < \sigma_2^2$,

iv. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) > \alpha_0$ when $\sigma_1^2 > \sigma_2^2$,

v. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) \to 0$ as $\sigma_1^2/\sigma_2^2 \to 0$,

vi. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) \to 1$ as $\sigma_1^2/\sigma_2^2 \to \infty$.

The test δ has level α_0 and is unbiased. The p-value when $V = v$ is observed equals $1 - G_{m-1,n-1}(v)$.

Proof The power function is the probability of rejecting H_0, i.e., the probability that $V \ge c$. Let V^* be as defined in Eq. (9.7.5) so that V^* has the F distribution with $m - 1$ and $n - 1$ degrees of freedom. Then

$$\begin{aligned}\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) = \Pr(V \ge c) &= \Pr\left(\frac{\sigma_1^2}{\sigma_2^2}V^* \ge c\right) = \Pr\left(V^* \ge \frac{\sigma_2^2}{\sigma_1^2}c\right)\\ &= 1 - G_{m-1,n-1}\left(\frac{\sigma_2^2}{\sigma_1^2}c\right),\end{aligned} \tag{9.7.6}$$

which proves property (i). Property (ii) follows from Theorem 9.7.3. For property (iii), let $\sigma_1^2 < \sigma_2^2$ in Eq. (9.7.6). Since $(\sigma_2^2/\sigma_1^2)c > c$, the expression on the far right of (9.7.6) is less than $1 - G_{m-1,n-1}(c) = \alpha_0$. Similarly, if $\sigma_1^2 > \sigma_2^2$, the expression on the far right of (9.7.6) is greater than $1 - G_{m-1,n-1}(c) = \alpha_0$, proving property (iv). Properties (v) and (vi) follow from property (i) and elementary properties of c.d.f.'s, namely, Property 3.3.2. The fact that δ has level α_0 follows from properties (ii) and (iii). The fact that δ is unbiased follows from properties (ii) and (iv). Finally, the p-value is the smallest α_0 such that we would reject H_0 at level α_0 if $V = v$ were observed. We reject H_0 at level α_0 if and only if $v \ge G_{m-1,n-1}^{-1}(1 - \alpha_0)$, which is equivalent to $\alpha_0 \ge 1 - G_{m-1,n-1}(v)$. Hence, $1 - G_{m-1,n-1}(v)$ is the smallest α_0 such that we would reject H_0. ■

Example 9.7.3 Performing an F Test. Suppose that six observations $X_1, \ldots, X_6$ are selected at random from a normal distribution for which both the mean μ_1 and the variance σ_1^2 are unknown, and it is found that $S_X^2 = 30$. Suppose also that 21 observations, $Y_1, \ldots, Y_{21}$, are selected at random from another normal distribution for which both the mean μ_2 and the variance σ_2^2 are unknown, and that it is found that $S_Y^2 = 40$. We shall carry out an F test of the hypotheses (9.7.3).

In this example, $m = 6$ and $n = 21$. Therefore, when H_0 is true, the statistic V defined by Eq. (9.7.4) will have the F distribution with 5 and 20 degrees of freedom. It follows from Eq. (9.7.4) that the value of V for the given samples is

$$V = \frac{30/5}{40/20} = 3.$$

It is found from the tables given at the end of this book that the 0.95 quantile of the F distribution with 5 and 20 degrees of freedom is 2.71, and the 0.975 quantile of that distribution is 3.29. Hence, the tail area corresponding to the value $V = 3$ is less than 0.05 and greater than 0.025. The hypothesis H_0 that $\sigma_1^2 \leq \sigma_2^2$ would therefore be rejected at the level of significance $\alpha_0 = 0.05$, and H_0 would not be rejected at the level of significance $\alpha_0 = 0.025$. (Using a computer program to evaluate the c.d.f. of an F distribution provides the p-value equal to 0.035.) Finally, suppose that it is important to reject H_0 if σ_1^2 is three times as large as σ_2^2. We would then want the power function to be high when $\sigma_1^2 = 3\sigma_2^2$. We use a computer program to compute

$$1 - F_{5,20}\left(2.71 \times \frac{1}{3}\right) = 0.498.$$

Even if σ_1^2 is three times as large as σ_2^2, the level 0.05 test only has about a 50 percent chance of rejecting H_0. ◀

Two-Sided Alternative

Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \sigma_1^2 = \sigma_2^2, \\ H_1&: \quad \sigma_1^2 \neq \sigma_2^2. \end{aligned} \tag{9.7.7}$$

It would make sense to reject H_0 if either $V \leq c_1$ or $V \geq c_2$, where V is defined in Eq. (9.7.4) and c_1 and c_2 are constants such that $\Pr(V \leq c_1) + \Pr(V \geq c_2) = \alpha_0$ when $\sigma_1^2 = \sigma_2^2$. The most convenient choice of c_1 and c_2 is the one that makes $\Pr(V \leq c_1) = \Pr(V \geq c_2) = \alpha_0/2$. That is, choose c_1 and c_2 to be the $\alpha_0/2$ and $1 - \alpha_0/2$ quantiles of the appropriate F distribution.

Example 9.7.4 Rain from Seeded Clouds. In Example 9.6.2, we compared the means of log-rainfalls from seeded and unseeded clouds under the assumption that the two variances were the same. We can now test the null hypothesis that the two variances are the same against the alternative hypothesis that the two variances are different at level of significance $\alpha_0 = 0.05$. Using the statistics given in Example 9.6.2, the value of V is $63.96/67.39 = 0.9491$, since $m = n$. We need to compare this to the 0.025 and 0.975 quantiles of the F distribution with 25 and 25 degrees of freedom. Since our table of F distribution quantiles does not have rows or columns for 25 degrees of freedom, we can either interpolate between 20 and 30 degrees of freedom or use a computer program to compute these quantiles. The quantiles are 0.4484 and 2.2303. Since V is between these two numbers, we would not reject the null hypothesis at level $\alpha_0 = 0.05$. ◀

When $m \neq n$, the two-sided F test constructed above is not unbiased. (See Exercise 19.) Also, if $m \neq n$, it is not possible to write the two-sided F test described above in the form “reject the null hypothesis if $T \geq c$” using the same statistic T for each significance level α_0. Nevertheless, we can still compute the smallest α_0 such that the two-sided F test with level of significance α_0 would reject H_0. The proof of the following result is left to Exercise 15 in this section.

Theorem 9.7.5 P-Value of Equal-Tailed Two-Sided F Test. Let V be as defined in (9.7.4). Suppose that we wish to test the hypotheses (9.7.7). Let δ_{α_0} be the equal-tailed two-sided F test that rejects H_0 when $V \le c_1$ or $V \ge c_2$, where c_1 and c_2 are, respectively, the $\alpha_0/2$ and $1-\alpha_0/2$ quantiles of the appropriate F distribution. Then the smallest α_0 such that δ_{α_0} rejects H_0 when $V = v$ is observed is

$$2\min\{1 - G_{m-1,\,n-1}(v),\, G_{m-1,\,n-1}(v)\}. \tag{9.7.8}$$

The F Test as a Likelihood Ratio Test

Next, we shall show that the F test for hypotheses (9.7.3) is a likelihood ratio test. After the values $x_1, \dots, x_m$ and $y_1, \dots, y_n$ in the two samples have been observed, the likelihood function $g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ is

$$g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = f_m(\boldsymbol{x}|\mu_1, \sigma_1^2) f_n(\boldsymbol{y}|\mu_2, \sigma_2^2).$$

Here, both $f_m(\boldsymbol{x}|\mu_1, \sigma_1^2)$ and $f_n(\boldsymbol{y}|\mu_2, \sigma_2^2)$ have the general form given in Eq. (9.5.9). For the hypotheses in (9.7.3), Ω_0 contains all parameters $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ with $\sigma_1^2 \le \sigma_2^2$, and Ω_1 contains all θ with $\sigma_1^2 > \sigma_2^2$. The likelihood ratio statistic is

$$\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sup_{\{(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2):\sigma_1^2\le\sigma_2^2\}} g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)}{\sup_{(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)} g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)}. \tag{9.7.9}$$

The likelihood ratio test then specifies that H_0 should be rejected if $\Lambda(\boldsymbol{x}, \boldsymbol{y}) \le k$, where k is typically chosen to make the test have a desired level α_0.

To facilitate the maximizations in (9.7.9), let

$$s_x^2 = \sum_{i=1}^{m}(x_i - \overline{x}_m)^2, \text{ and } s_y^2 = \sum_{i=1}^{n}(y_i - \overline{y}_n)^2.$$

Then we can write

$$g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$$
$$= \frac{1}{(2\pi)^{(m+n)/2}\sigma_1^m\sigma_2^n} \exp\left(-\frac{1}{2\sigma_1^2}\left[n(\overline{x}_m - \mu_1)^2 + s_x^2\right] - \frac{1}{2\sigma_2^2}\left[n(\overline{y}_n - \mu_2)^2 + s_y^2\right]\right).$$

For both the numerator and denominator of (9.7.9), we need $\mu_1 = \overline{x}_m$ and $\mu_2 = \overline{y}_n$ in order to maximize the likelihood. If $s_x^2/m \le s_y^2/n$, then the numerator is maximized at $\sigma_1^2 = s_x^2/m$ and $\sigma_2^2 = s_y^2/n$. These values also maximize the denominator. Hence, $\Lambda(\boldsymbol{x}, \boldsymbol{y}) = 1$ if $s_x^2/m \le s_y^2/n$. For the other case (the numerator when $s_x^2/m > s_y^2/n$), it is straightforward to show that $\sigma_1^2 = \sigma_2^2$ is required in order to achieve the maximum. In these cases, the maximum occurs when

$$\sigma_1^2 = \sigma_2^2 = \frac{s_x^2 + s_y^2}{m+n}.$$

Substituting all of these values into (9.7.9) yields

$$\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 1 & \text{if } s_x^2/m \le s_y^2/n, \\ dw^{m/2}(1-w)^{n/2} & \text{if } s_x^2/m > s_y^2/m, \end{cases}$$

where

$$w = \frac{s_x^2}{s_x^2 + s_y^2}, \quad \text{and} \quad d = \frac{(m+n)^{(m+n)/2}}{m^{m/2} n^{n/2}}.$$

Note that $s_x^2/m \le s_y^2/n$ if and only if $w \le m/(m+n)$. Next, use the fact that the function $h(w) = w^{m/2}(1-w)^{n/2}$ is decreasing for $m/(m+n) < w < 1$. Finally, note that $h(m/[m+n]) = 1/d$. For $k < 1$, it follows that $\Lambda(\boldsymbol{x}, \boldsymbol{y}) \le k$ if and only if $w \ge k'$ for some other constant k'. This, in turn, is equivalent to $s_x^2/s_y^2 \ge k''$. Since s_x^2/s_y^2 is a positive constant times the observed value of V, the likelihood ratio test rejects H_0 when V is large. This is the same as the F test.

One can easily adapt the above argument for the case in which the inequalities are reversed in the hypotheses. When the hypotheses are (9.7.7), that is, the alternative is two-sided, one can show (see Exercise 16) that the size α_0 likelihood ratio test will reject H_0 if either $V \le c_1$ or $V \ge c_2$. Unfortunately, it is usually tedious to compute the necessary values c_1 and c_2. For this reason, people often abandon the strict likelihood ratio criterion in this case and simply let c_1 and c_2 be the $\alpha_0/2$ and $1 - \alpha_0/2$ quantiles of the appropriate F distribution.

Derivation of the p.d.f. of an F distribution

Since the random variables Y and W in Definition 9.7.1 are independent, their joint p.d.f. $g(y, w)$ is the product of their individual p.d.f.'s. Furthermore, since both Y and W have χ^2 distributions, it follows from the p.d.f. of the χ^2 distribution, as given in Eq. 8.2.1, that $g(y, w)$ has the following form, for $y > 0$ and $w > 0$:

$$g(y, w) = c y^{(m/2)-1} w^{(n/2)-1} e^{-(y+w)/2}, \tag{9.7.10}$$

where

$$c = \frac{1}{2^{(m+n)/2}\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)}. \tag{9.7.11}$$

We shall now change variables from Y and W to X and W, where X is defined by Eq. (9.7.1). The joint p.d.f. $h(x, w)$ of X and W is obtained by first replacing y in Eq. (9.7.10) with its expression in terms of x and w and then multiplying the result by $|\partial y/\partial x|$. It follows from Eq. (9.7.1) that $y = (m/n)xw$ and $\partial y/\partial x = (m/n)w$. Hence, the joint p.d.f. $h(x, w)$ has the following form, for $x > 0$ and $w > 0$:

$$h(x, w) = c\left(\frac{m}{n}\right)^{m/2} x^{(m/2)-1} w^{[(m+n)/2]-1} \exp\left[-\frac{1}{2}\left(\frac{m}{n}x + 1\right)w\right]. \tag{9.7.12}$$

Here, the constant c is again given by Eq. (9.7.11).

The marginal p.d.f $f(x)$ of X can be obtained for each value of $x > 0$ from the relation

$$f(x) = \int_0^\infty h(x, w)\, dw. \tag{9.7.13}$$

It follows from Theorem 5.7.3 that

$$\int_0^\infty w^{[(m+n)/2]-1} \exp\left[-\frac{1}{2}\left(\frac{m}{n}x + 1\right)w\right] dw = \frac{\Gamma\left[\frac{1}{2}(m+n)\right]}{\left[\frac{1}{2}\left(\frac{m}{n}x + 1\right)\right]^{(m+n)/2}}. \tag{9.7.14}$$

From Eqs. (9.7.11) to (9.7.14), we can conclude that the p.d.f $f(x)$ has the form given in Eq. (9.7.2).

Summary

If Y and W are independent with Y having the χ^2 distribution with m degrees of freedom and W having the χ^2 distribution with n degrees of freedom, then $(Y/m)/(W/n)$ has the F distribution with m and n degrees of freedom. Suppose that we observe two independent random samples from two normal distributions with possibly different variances. The ratio V of the usual unbiased estimators of the two variances will have an F distribution when the two variances are equal. Tests of hypotheses about the two variances can be constructed by comparing V to various quantiles of F distributions.

Exercises

1. Consider again the situation described in Exercise 11 of Sec. 9.6. Test the null hypothesis that the variance of the fusion time for subjects who saw a picture of the object is no smaller than the variance for subjects who did see a picture. The alternative hypothesis is that the variance for subjects who saw a picture is smaller than the variance for subjects who did not see a picture. Use a level of significance of 0.05.

2. Suppose that a random variable X has the F distribution with three and eight degrees of freedom. Determine the value of c such that $\Pr(X > c) = 0.975$.

3. Suppose that a random variable X has the F distribution with one and eight degrees of freedom. Use the table of the t distribution to determine the value of c such that $\Pr(X > c) = 0.3$.

4. Suppose that a random variable X has the F distribution with m and n degrees of freedom ($n > 2$). Show that $E(X) = n/(n-2)$. *Hint:* Find the value of $E(1/Z)$, where Z has the χ^2 distribution with n degrees of freedom.

5. What is the value of the median of the F distribution with m and n degrees of freedom when $m = n$?

6. Suppose that a random variable X has the F distribution with m and n degrees of freedom. Show that the random variable $mX/(mX + n)$ has the beta distribution with parameters $\alpha = m/2$ and $\beta = n/2$.

7. Consider two different normal distributions for which both the means μ_1 and μ_2 and the variances σ_1^2 and σ_2^2 are unknown, and suppose that it is desired to test the following hypotheses:

$$H_0:\ \sigma_1^2 \le \sigma_2^2,$$
$$H_1:\ \sigma_1^2 > \sigma_2^2.$$

Suppose further that a random sample consisting of 16 observations for the first normal distribution yields the values $\sum_{i=1}^{16} X_i = 84$ and $\sum_{i=1}^{16} X_i^2 = 563$, and an independent random sample consisting of 10 observations from the second normal distribution yields the values $\sum_{i=1}^{10} Y_i = 18$ and $\sum_{i=1}^{10} Y_i^2 = 72$.

a. What are the M.L.E.'s of σ_1^2 and σ_2^2?

b. If an F test is carried out at the level of significance 0.05, is the hypothesis H_0 rejected or not?

8. Consider again the conditions of Exercise 7, but suppose now that it is desired to test the following hypotheses:

$$H_0:\ \sigma_1^2 \le 3\sigma_2^2,$$
$$H_1:\ \sigma_1^2 > 3\sigma_2^2.$$

Describe how to carry out an F test of these hypotheses.

9. Consider again the conditions of Exercise 7, but suppose now that it is desired to test the following hypotheses:

$$H_0:\ \sigma_1^2 = \sigma_2^2,$$
$$H_1:\ \sigma_1^2 \ne \sigma_2^2.$$

Suppose also that the statistic V is defined by Eq. (9.7.4), and it is desired to reject H_0 if either $V \le c_1$ or $V \ge c_2$, where the constants c_1 and c_2 are chosen so that when H_0 is true, $\Pr(V \le c_1) = \Pr(V \ge c_2) = 0.025$. Determine the values of c_1 and c_2 when $m = 16$ and $n = 10$, as in Exercise 7.

10. Suppose that a random sample consisting of 16 observations is available from the normal distribution for which both the mean μ_1 and the variance σ_1^2 are unknown, and an independent random sample consisting of 10 observations is available from the normal distribution for which both the mean μ_2 and the variance σ_2^2 are also unknown.

For each constant $r > 0$, construct a test of the following hypotheses at the level of significance 0.05:

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = r, \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq r.$$

11. Consider again the conditions of Exercise 10. Use the results of that exercise to construct a confidence interval for σ_1^2/σ_2^2 with confidence coefficient 0.95.

12. Suppose that a random variable Y has the χ^2 distribution with m_0 degrees of freedom, and let c be a constant such that $\Pr(Y > c) = 0.05$. Explain why, in the table of 0.95 quantile of the F distribution, the entry for $m = m_0$ and $n = \infty$ will be equal to c/m_0.

13. The final column in the table of the 0.95 quantile of the F distribution contains values for which $m = \infty$. Explain how to derive the entries in this column from a table of the χ^2 distribution.

14. Consider again the conditions of Exercise 7. Find the power function of the F test when $\sigma_1^2 = 2\sigma_2^2$.

15. Prove Theorem 9.7.5. Also, compute the p-value for Example 9.7.4 using the formula in Eq. (9.7.8).

16. Let V be as defined in Eq. (9.7.4). We wish to determine the size α_0 likelihood ratio test of the hypotheses (9.7.7). Prove that the likelihood ratio test will reject H_0 if either $V \le c_1$ or $V \ge c_2$, where $\Pr(V \le c_1) + \Pr(V \ge c_2) = \alpha_0$ when $\sigma_1^2 = \sigma_2^2$.

17. Prove that the test found in Exercise 9 is *not* a likelihood ratio test.

18. Let δ be the two-sided F test that rejects H_0 in (9.7.3) when either $V \le c_1$ or $V \ge c_2$ with $c_1 < c_2$. Prove that the power function of δ is

$$\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) = G_{m-1,n-1}\left(\frac{\sigma_2^2}{\sigma_1^2}c_1\right) + 1 - G_{m-1,n-1}\left(\frac{\sigma_2^2}{\sigma_1^2}c_2\right).$$

19. Suppose that $X_1, \dots, X_{11}$ form a random sample from the normal distribution with unknown mean μ_1 and unknown variance σ_1^2. Suppose also that $Y_1, \dots, Y_{21}$ form an independent random sample from the normal distribution with unknown mean μ_2 and unknown variance σ_2^2. Suppose that we wish to test the hypotheses in Eq. (9.7.7). Let δ be the equal-tailed two-sided F test with level of significance $\alpha_0 = 0.5$.

a. Compute the power function of δ when $\sigma_1^2 = 1.01\sigma_2^2$.

b. Compute the power function of δ when $\sigma_1^2 = \sigma_2^2/1.01$.

c. Show that δ is not an unbiased test. (You will probably need computer software that computes the function $G_{m-1,n-1}$. And try to minimize the amount of rounding you do.)

⋆ 9.8 Bayes Test Procedures

Here we summarize how one tests hypotheses from the Bayesian perspective. The general idea is to choose the action (reject H_0 or not) that leads to the smaller posterior expected loss. We assume that the loss of making an incorrect decision is larger than the loss of making a correct decision. Many of the Bayes test procedures have the same forms as the tests we have already seen, but their interpretations are different.

Simple Null and Alternative Hypotheses

Example 9.8.1 Service Times in a Queue. In Example 9.2.1, a manager was trying to decide which of two joint distributions better describes customer service times. She was comparing the two joint p.d.f.'s f_1 and f_0 in Eqs. (9.2.1) and (9.2.2), respectively. Suppose that there are costs involved in making a bad choice. For example, if she chooses a joint distribution that models the service times as shorter than they really tend to be, there may be a cost due to customers becoming frustrated and taking their business elsewhere. On the other hand, if she chooses a joint distribution that models the service times as longer than they really tend to be, there may be a cost due to hiring additional unnecessary servers. How should the manager weigh these costs together with available evidence about how long she believes service times tend to be in order to choose between the two joint distributions? ◀

Consider a general problem in which the parameter space consists of two values $\Omega = \{\theta_0, \theta_1\}$. If $\theta = \theta_i$ (for $i = 0, 1$), let $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f_i(\boldsymbol{x})$. Suppose that it is desired to test the following simple hypotheses:

$$\begin{aligned} H_0&: \quad \theta = \theta_0, \\ H_1&: \quad \theta = \theta_1. \end{aligned} \tag{9.8.1}$$

We shall let d_0 denote the decision not to reject the hypothesis H_0 and let d_1 denote the decision to reject H_0. Also, we shall assume that the losses resulting from choosing an incorrect decision are as follows: If decision d_1 is chosen when H_0 is actually the true hypothesis (type I error), then the loss is w_0 units; if decision d_0 is chosen when H_1 is actually the true hypothesis (type II error), then the loss is w_1 units. If the decision d_0 is chosen when H_0 is the true hypothesis or if the decision d_1 is chosen when H_1 is the true hypothesis, then the correct decision has been made and the loss is 0. Thus, for $i = 0, 1$ and $j = 0, 1$, the loss $L(\theta_i, d_j)$ that occurs when θ_i is the true value of θ and the decision d_j is chosen is given by the following table:

	d_0	d_1
θ_0	0	w_0
θ_1	w_1	0

(9.8.2)

Next, suppose that the prior probability that H_0 is true is ξ_0, and the prior probability that H_1 is true is $\xi_1 = 1 - \xi_0$. Then the expected loss $r(\delta)$ of each test procedure δ will be

$$r(\delta) = \xi_0 E(\text{Loss } |\theta = \theta_0) + \xi_1 E(\text{Loss}|\theta = \theta_1). \tag{9.8.3}$$

If $\alpha(\delta)$ and $\beta(\delta)$ again denote the probabilities of the two types of errors for the procedure δ, and if the table of losses just given is used, it follows that

$$\begin{aligned} E(\text{Loss}|\theta = \theta_0) &= w_0 \Pr(\text{Choosing } d_1|\theta = \theta_0) = w_0\alpha(\delta), \\ E(\text{Loss}|\theta = \theta_1) &= w_1 \Pr(\text{Choosing } d_0|\theta = \theta_1) = w_1\beta(\delta). \end{aligned} \tag{9.8.4}$$

Hence,

$$r(\delta) = \xi_0 w_0 \alpha(\delta) + \xi_1 w_1 \beta(\delta). \tag{9.8.5}$$

A procedure δ for which this expected loss $r(\delta)$ is minimized is called a *Bayes test procedure*.

Since $r(\delta)$ is simply a linear combination of the form $a\alpha(\delta) + b\beta(\delta)$ with $a = \xi_0 w_0$ and $b = \xi_1 w_1$, a Bayes test procedure can immediately be determined from Theorem 9.2.1. Thus, a Bayes procedure will not reject H_0 whenever $\xi_0 w_0 f_0(\boldsymbol{x}) > \xi_1 w_1 f_1(\boldsymbol{x})$ and will reject H_0 whenever $\xi_0 w_0 f_0(\boldsymbol{x}) < \xi_1 w_1 f_1(\boldsymbol{x})$. We can either reject H_0 or not if $\xi_0 w_0 f_0(\boldsymbol{x}) = \xi_1 w_1 f_1(\boldsymbol{x})$. For simplicity, in the remainder of this section, we shall assume that H_0 is rejected whenever $\xi_0 w_0 f_0(\boldsymbol{x}) = \xi_1 w_1 f_1(\boldsymbol{x})$.

Note: Bayes Test Depends Only on the Ratio of Costs. Notice that choosing δ to minimize $r(\delta)$ in Eq. (9.8.5) is not affected if we multiply w_0 and w_1 by the same positive constant, such as $1/w_0$. That is, the Bayes test δ is also the test that minimizes

$$r^*(\delta) = \xi_0 \alpha(\delta) + \xi_1 \frac{w_1}{w_0} \beta(\delta).$$

So, a decision maker does not need to choose both of the two costs of error, but rather just the ratio of the two costs. One can think of choosing the ratio of costs as a replacement for specifying a level of significance when selecting a test procedure.

Example 9.8.2 Service Times in a Queue. Suppose that the manager believes that each of the two models for service times is equally likely before observing any data so that $\xi_0 = \xi_1 = 1/2$. The model with joint p.d.f. f_1 predicts both extremely large service times and extremely small service times to be more likely than does the model with joint p.d.f. f_0. Suppose that the cost of modeling extremely large service times as being less likely than they really are is the same as the cost of modeling extremely large service times to be more likely than they really are. The ratio of the cost of type II error w_1 to the cost of type I error w_0 is then $w_1/w_0 = 1$. The Bayes test is then to choose d_1 (reject H_0) if $f_0(\boldsymbol{x}) < f_1(\boldsymbol{x})$. This is equivalent to $f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) > 1$. ◀

Tests Based on the Posterior Distribution

From the Bayesian viewpoint, it is more natural to base a test on the posterior distribution of θ rather than on the prior distribution and the probabilities of error as we did in the preceding discussion. Fortunately, the same test procedure arises regardless of how one derives it. For example, Exercise 5 in this section asks you to prove that the test derived by minimizing a linear combination of error probabilities is the same as what one would obtain by minimizing the posterior expected value of the loss. The same is true in general when the losses are bounded, but the proof is more difficult. For the remainder of this section, we shall take the more natural approach of trying to minimize the posterior expected value of the loss directly.

Return again to the general situation in which the null hypothesis is $H_0 : \theta \in \Omega_0$ and the alternative hypothesis is $H_1 : \theta \in \Omega_1$, where $\Omega_0 \cup \Omega_1$ is the entire parameter space. As we did above, we shall let d_0 denote the decision not to reject the null hypothesis H_0 and let d_1 denote the decision to reject H_0. As before, we shall assume that we incur a loss of w_0 by making decision d_1 when H_0 is actually true, and a loss of w_1 is incurred if we make decision d_0 when H_1 is true. (More realistic loss functions are available, but this simple type of loss will suffice for an introduction.) The loss function $L(\theta, d_i)$ can be summarized in the following table:

	d_0	d_1
If H_0 is true	0	w_0
If H_1 is true	w_1	0

(9.8.6)

We shall now take the approach outlined in Exercise 5. Suppose that $\xi(\theta|\boldsymbol{x})$ is the posterior p.d.f. for θ. Then the posterior expected loss $r(d_i|\boldsymbol{x})$ for choosing decision d_i $(i = 0, 1)$ is

$$r(d_i|\boldsymbol{x}) = \int L(\theta, d_i)\xi(\theta|\boldsymbol{x})\, d\theta.$$

We can write a simpler formula for this posterior expected loss for each of $i = 0, 1$:

$$r(d_0|\boldsymbol{x}) = \int_{\Omega_1} w_1\xi(\theta|\boldsymbol{x})\, d\theta = w_1[1 - \Pr(H_0 \text{ true}|\boldsymbol{x})],$$
$$r(d_1|\boldsymbol{x}) = \int_{\Omega_0} w_0\xi(\theta|\boldsymbol{x})\, d\theta = w_0 \Pr(H_0 \text{ true}|\boldsymbol{x}).$$

The Bayes test procedure is to choose the decision that has the smaller posterior expected loss, that is, choose d_0 if $r(d_0|\boldsymbol{x}) < r(d_1|\boldsymbol{x})$, choose d_1 if $r(d_0|\boldsymbol{x}) \ge r(d_1|\boldsymbol{x})$. Using the expressions above, it is easy to see that the inequality $r(d_0|\boldsymbol{x}) \ge r(d_1|\boldsymbol{x})$

(when to reject H_0) can be rewritten as

$$\Pr(H_0 \text{ true}|\boldsymbol{x}) \leq \frac{w_1}{w_0 + w_1}, \tag{9.8.7}$$

just as in part (c) of Exercise 5.

The test procedure that rejects H_0 when (9.8.7) holds is the Bayes test in all situations in which the loss function is given by the table in (9.8.6). This result holds whether or not the distributions have monotone likelihood ratio, and it even applies when the alternative is two-sided or when the parameter is discrete rather than continuous. Furthermore, the Bayes test produces the same result if one were to switch the names of H_0 and H_1, as well as the losses w_0 and w_1 and the names of the decisions d_0 and d_1. (See Exercise 11 in this section.)

Despite the generality of (9.8.7), it is instructive to examine what the procedure looks like in special cases that we have already encountered.

One-Sided Hypotheses

Suppose that the family of distributions has a monotone likelihood ratio and that the hypotheses are

$$\begin{aligned} H_0&: \quad \theta \leq \theta_0, \\ H_1&: \quad \theta > \theta_0. \end{aligned} \tag{9.8.8}$$

We shall prove next that the Bayes procedure that rejects H_0 when (9.8.7) holds is a one-sided test as in Theorem 9.3.1.

Theorem 9.8.1 Suppose that $f_n(\boldsymbol{x}|\theta)$ has a monotone likelihood ratio in the statistic $T = r(\boldsymbol{X})$. Let the hypotheses be as in Eq. (9.8.8), and assume that the loss function is of the form

	d_0	d_1
$\theta \leq \theta_0$	0	w_0
$\theta > \theta_0$	w_1	0

where $w_0, w_1 > 0$ are constants. Then a test procedure that minimizes the posterior expected loss is to reject H_0 when $T \geq c$ for some constant c (possibly infinite).

Proof According to Bayes' theorem for parameters and samples, (7.2.7), the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ can be expressed as

$$\xi(\theta|\boldsymbol{x}) = \frac{f_n(\boldsymbol{x}|\theta)\xi(\theta)}{\int_\Omega f_n(\boldsymbol{x}|\psi)\xi(\psi)\,d\psi}.$$

The ratio of the posterior expected loss from making decision d_0 to the posterior expected loss from making decision d_1 after observing $\boldsymbol{X} = \boldsymbol{x}$ is

$$\ell(\boldsymbol{x}) = \frac{\int_{\theta_0}^{\infty} w_1\xi(\theta|\boldsymbol{x})\,d\theta}{\int_{-\infty}^{\theta_0} w_0\xi(\psi|\boldsymbol{x})\,d\psi} = \frac{w_1\int_{\theta_0}^{\infty} f_n(\boldsymbol{x}|\theta)\xi(\theta)\,d\theta}{w_0\int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}|\psi)\xi(\psi)\,d\psi}. \tag{9.8.9}$$

What we need to prove is that $\ell(\boldsymbol{x}) \geq 1$ is equivalent to $T \geq c$. It suffices to show that $\ell(\boldsymbol{x})$ is a nondecreasing function in $T = r(\boldsymbol{x})$. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be two possible observations with the property that $r(\boldsymbol{x}_1) \leq r(\boldsymbol{x}_2)$. We want to prove that $\ell(\boldsymbol{x}_1) \leq \ell(\boldsymbol{x}_2)$.

We can write

$$\ell(\boldsymbol{x}_1) - \ell(\boldsymbol{x}_2) = \frac{w_1 \int_{\theta_0}^{\infty} f_n(\boldsymbol{x}_1|\theta)\xi(\theta)\, d\theta}{w_0 \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_1|\psi)\xi(\psi)\, d\psi} - \frac{w_1 \int_{\theta_0}^{\infty} f_n(\boldsymbol{x}_2|\theta)\xi(\theta)\, d\theta}{w_0 \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_2|\psi)\xi(\psi)\, d\psi}. \tag{9.8.10}$$

We can put the two fractions on the right side of Eq. (9.8.10) over the common denominator $w_0^2 \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_2|\psi)\xi(\psi)\, d\psi \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_1|\psi)\xi(\psi)\, d\psi$. The numerator of the resulting fraction is $w_0 w_1$ times

$$\begin{aligned}
&\int_{\theta_0}^{\infty} f_n(\boldsymbol{x}_1|\theta)\xi(\theta)\, d\theta \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_2|\psi)\xi(\psi)\, d\psi \\
&- \int_{\theta_0}^{\infty} f_n(\boldsymbol{x}_2|\theta)\xi(\theta)\, d\theta \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_1|\psi)\xi(\psi)\, d\psi.
\end{aligned} \tag{9.8.11}$$

We only need to show that (9.8.11) is at most 0. The difference in (9.8.11) can be written as the double integral

$$\int_{\theta_0}^{\infty} \int_{-\infty}^{\theta_0} \xi(\theta)\xi(\psi)[f_n(\boldsymbol{x}_1|\theta) f_n(\boldsymbol{x}_2|\psi) - f_n(\boldsymbol{x}_2|\theta) f_n(\boldsymbol{x}_1|\psi)]\, d\psi\, d\theta. \tag{9.8.12}$$

Notice that for all θ and ψ in this double integral, $\theta \ge \theta_0 \ge \psi$. Since $r(\boldsymbol{x}_1) \le r(\boldsymbol{x}_2)$, monotone likelihood ratio implies that

$$\frac{f_n(\boldsymbol{x}_1|\theta)}{f_n(\boldsymbol{x}_1|\psi)} - \frac{f_n(\boldsymbol{x}_2|\theta)}{f_n(\boldsymbol{x}_2|\psi)} \le 0.$$

If one multiplies both sides of this last expression by the product of the two denominators, the result is

$$f_n(\boldsymbol{x}_1|\theta) f_n(\boldsymbol{x}_2|\psi) - f_n(\boldsymbol{x}_2|\theta) f_n(\boldsymbol{x}_1|\psi) \le 0. \tag{9.8.13}$$

Notice that the left side of Eq. (9.8.13) appears inside the square brackets in the integrand of (9.8.12). Since this is nonpositive, it implies that (9.8.12) is at most 0, and so (9.8.11) is at most 0. ■

Example 9.8.3

Calorie Counts on Food Labels. In Example 7.3.10 on page 400, we were interested in the percentage differences between the observed and advertised calorie counts for nationally prepared foods. We modeled the differences $X_1, \ldots, X_{20}$ as normal random variables with mean θ and variance 100. The prior for θ was a normal distribution with mean 0 and variance 60. The family of normal distributions has a monotone likelihood ratio in the statistic $\overline{X}_{20} = \frac{1}{20}\sum_{i=1}^{20} X_i$. The posterior distribution of θ is the normal distribution with mean

$$\mu_1 = \frac{100 \times 0 + 20 \times 60 \times \overline{X}_{20}}{100 + 20 \times 60} = 0.923\overline{X}_{20}$$

and variance $v_1^2 = 4.62$. Suppose that we wish to test the null hypothesis $H_0: \theta \le 0$ versus the alternative $H_1: \theta > 0$. The posterior probability that H_0 is true is

$$\Pr(\theta \le 0|\overline{X}_{20}) = \Phi\left(\frac{0 - \mu_1}{v_1}\right) = \Phi(-0.429\overline{X}_{20}).$$

The Bayes test will reject H_0 if this probability is at most $w_1/(w_0 + w_1)$. Since Φ is a strictly increasing function, $\Phi(-0.429\overline{X}_{20}) \le w_1/(w_0 + w_1)$ if and only if $\overline{X}_{20} \ge -\Phi^{-1}(w_1/(w_0 + w_1))/0.492$. This is in the form of a one-sided test. ◄

Two-Sided Alternatives

On page 571, we argued that the hypotheses

$$\begin{aligned} H_0&: \quad \theta = \theta_0, \\ H_1&: \quad \theta \neq \theta_0 \end{aligned} \tag{9.8.14}$$

might be a useful surrogate for the null hypothesis that θ is close to θ_0 against the alternative that it is not close. If the prior distribution of θ is continuous, then the posterior distribution will usually be continuous as well. In such cases, the posterior probability that H_0 is true will be 0, and H_0 would be rejected without having to refer to the data. If one believed that $\theta = \theta_0$ with positive probability, one should use a prior distribution that is not continuous, but we shall not take that approach here. (See a more advanced text, such as Schervish, 1995, section 4.2, for treatment of that approach.) Instead, we can calculate the posterior probability that θ is close to θ_0. If this probability is too small, we can reject the null hypothesis that θ is close to θ_0. To be specific, let $d > 0$, and consider the hypotheses

$$\begin{aligned} H_0&: \quad |\theta - \theta_0| \le d, \\ H_1&: \quad |\theta - \theta_0| > d. \end{aligned} \tag{9.8.15}$$

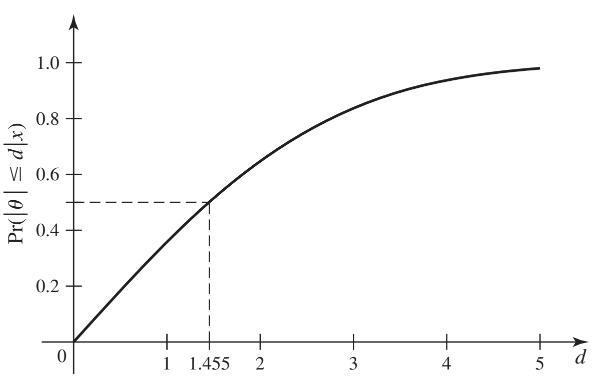
Many experimenters might choose to test the hypotheses in (9.8.14) rather than those in (9.8.15) because they are not ready to specify a particular value of d. In such cases, one could calculate the posterior probability of $|\theta - \theta_0| \le d$ for all d and draw a little plot.

Example 9.8.4 Calorie Counts on Food Labels. Suppose that we wish to test the hypotheses (9.8.15) with $\theta_0 = 0$ in the situation described in Example 9.8.3. In Example 7.3.10, we found that the posterior distribution of θ was the normal distribution with mean 0.1154 and variance 4.62. We can easily calculate

$$\Pr(|\theta - 0| \le d|\boldsymbol{x}) = \Pr(-d \le \theta \le d|\boldsymbol{x}) = \Phi\left(\frac{d - 0.1154}{4.62^{1/2}}\right) - \Phi\left(\frac{-d - 0.1154}{4.62^{1/2}}\right),$$

for every value of d that we want. Figure 9.16 shows a plot of the posterior probability that $|\theta|$ is at most d for all values of d between 0 and 5. In particular, we see that $\Pr(|\theta| \le 5|\boldsymbol{x})$ is very close to 1. If 5 percent is considered a small discrepancy, then we can be pretty sure that $|\theta|$ is small. On the other hand, $\Pr(|\theta| \ge 1|\boldsymbol{x})$ is greater than 0.6. If 1 percent is considered large, then there is a substantial chance that $|\theta|$ is large. ◀

Figure 9.16 Plot of $\Pr(|\theta| \le d|\boldsymbol{x})$ against d for Example 9.8.4. The dotted lines indicate that the median of the posterior distribution of $|\theta|$ is 1.455.



Note: What Counts as a Meaningful Difference?. The method illustrated in Example 9.8.4 raises a useful point. In order to complete the test procedure, we need to decide what counts as a meaningful difference between θ and θ_0. Otherwise, we cannot say whether or not the probability is large that a meaningful difference exists. Forcing experimenters to think about what counts as a meaningful difference is a good idea. Testing the hypotheses (9.8.14) at a fixed level, such as 0.05, does not require anyone to think about what counts as a meaningful difference. Indeed, if an experimenter did bother to decide what counted as a meaningful difference, it is not clear how to make use of that information in choosing a significance level at which to test the hypotheses in (9.8.14).

Testing the Mean of a Normal Distribution with Unknown Variance

In Sec. 8.6, we considered the case in which a random sample is drawn from a normal distribution with unknown mean and variance. We introduced a family of conjugate prior distributions and found that the posterior distribution of a linear function of the mean μ is a t distribution. If we wish to test the null hypothesis that μ lies in an interval using (9.8.7) as the condition for rejecting the null hypothesis, then we only need a table or computer program to calculate the c.d.f. of an arbitrary t distribution. Most statistical software packages allow calculation of the c.d.f. and the quantile function of an arbitrary t distribution, and hence we can perform Bayes tests of null hypotheses of the form $\mu \le \mu_0$, $\mu \ge \mu_0$, or $d_1 \le \mu \le d_2$.

Example 9.8.5 Pesticide Residue on Celery. Sharpe and Van Middelem (1955) describe an experiment in which $n = 77$ samples of parathion residue were measured on celery after the vegetable had been taken from fields sprayed with parathion. Figure 9.17 shows a histogram of the observations. (Each concentration Z in parts per million was transformed to $X = 100(Z - 0.7)$ for ease of recording.) Suppose that we model the X values as normal with mean μ and variance σ^2. We will use an improper prior for μ and σ^2. The sample average is $\overline{x}_n = 50.23$, and

$$s_n^2 = \sum_{i=1}^{77} (x_i - \overline{x}_{77})^2 = 34106.$$

As we saw in Eq. (8.6.21), this means that the posterior distribution of

$$\frac{n^{1/2}(\mu - \overline{x}_n)}{(s_n^2/(n-1))^{1/2}} = \frac{77^{1/2}(\mu - 50.23)}{(34106/76)^{1/2}} = 0.4142\mu - 20.81$$

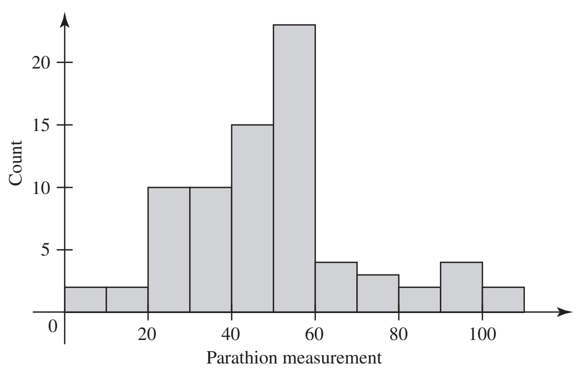


Figure 9.17 Histogram of parathion measurements on 77 celery samples.

is the t distribution with 76 degrees of freedom. Suppose that we are interested in testing the null hypothesis $H_0 : \mu \ge 55$ against the alternative $H_1 : \mu < 55$. Suppose that our losses are described by (9.8.6). Then we should reject H_0 if its posterior probability is at most $\alpha_0 = w_1/(w_0 + w_1)$. If we let T_{n-1} stand for the c.d.f. of the t distribution with $n-1$ degrees of freedom, we can write this probability as

$$\Pr(\mu \ge 55|\boldsymbol{x}) = \Pr\left(\frac{n^{1/2}(\mu - \overline{x}_n)}{(s_n^2/(n-1))^{1/2}} \ge \frac{n^{1/2}(55 - \overline{x}_n)}{(s_n^2/(n-1))^{1/2}} \middle| \boldsymbol{x} \right)$$
$$= 1 - T_{n-1}\left(\frac{n^{1/2}(55 - \overline{x}_n)}{(s_n^2/(n-1))^{1/2}} \right). \quad (9.8.16)$$

Simple manipulation shows that this last probability is at most α_0 if and only if $U \le T_{n-1}^{-1}(1-\alpha_0)$, where U is the random variable in Eq. (9.5.2) that was used to define the t test. Indeed, the level α_0 t test of H_0 versus H_1 is precisely to reject H_0 if $U \le T_{n-1}^{-1}(1-\alpha_0)$. For the data in this example, the probability in Eq. (9.8.16) is $1 - T_{76}(1.974) = 0.026$. ◄

Note: Look at Your Data. The histogram in Fig. 9.17 has a strange feature. Can you specify what it is? If you take a course in data analysis, you will probably learn some methods for dealing with data having features like this.

Note: Bayes Tests for One-Sided Nulls with Improper Priors Are t Tests. In Example 9.8.5, we saw that the Bayes test for one-sided hypotheses was the level α_0 t test for the same hypotheses where $\alpha_0 = w_1/(w_0 + w_1)$. This holds in general for normal data with improper priors. It also follows that the p-values in these cases must be the same as the posterior probabilities that the null hypotheses are true. (See Exercise 7 in this section.)

Comparing the Means of Two Normal Distributions

Next, consider the case in which we shall observe two independent normal random samples with common variance σ^2: $X_1, \dots, X_m$ with mean μ_1 and $Y_1, \dots, Y_n$ with mean μ_2. In order to use the Bayesian approach, we need the posterior distribution of $\mu_1 - \mu_2$. We could introduce a family of conjugate prior distributions for the three parameters μ_1, μ_2, and $\tau = 1/\sigma^2$, and then proceed as we did in Sec. 8.6. For simplicity, we shall only handle the case of improper priors in this section, although there are proper conjugate priors that will lead to more general results. The usual improper prior for each parameter μ_1 and μ_2 is the constant function 1, and the usual improper prior for τ is $1/\tau$ for $\tau > 0$. If we combine these as if the parameters were independent, the improper prior p.d.f. would be $\xi(\mu_1, \mu_2, \tau) = 1/\tau$ for $\tau > 0$. We can now find the posterior joint distribution of the parameters.

Theorem 9.8.2 Suppose that $X_1, \dots, X_m$ form a random sample from a normal distribution with mean μ_1 and precision τ while $Y_1, \dots, Y_n$ form a random sample from a normal distribution with mean μ_2 and precision τ. Suppose that the parameters have the improper prior with "p.d.f." $\xi(\mu_1, \mu_2, \tau) = 1/\tau$ for $\tau > 0$. The posterior distribution of

$$(m+n-2)^{1/2} \frac{\mu_1 - \mu_2 - (\overline{x}_m - \overline{y}_n)}{\left(\frac{1}{m} + \frac{1}{n} \right)^{1/2} (s_x^2 + s_y^2)^{1/2}} \quad (9.8.17)$$

is the t distribution with $m+n-2$ degrees of freedom, where s_x^2 and s_y^2 are the observed values of S_X^2 and S_Y^2, respectively. ■

The proof of Theorem 9.8.2 is left as Exercise 8 because it is very similar to results proven in Sec. 8.6.

For testing the hypotheses

$$\begin{aligned} H_0&: \quad \mu_1 - \mu_2 \le 0, \\ H_1&: \quad \mu_1 - \mu_2 > 0, \end{aligned}$$

we need the posterior probability that $\mu_1 - \mu_2 \le 0$, which is easily obtained from the posterior distribution. Using the same idea as in Eq. (9.8.16), we can write $\Pr(\mu_1 - \mu_2 \le 0 | \boldsymbol{x}, \boldsymbol{y})$ as the probability that the random variable in (9.8.17) is at most $-u$, where u is the observed value of the random variable U in Eq. (9.6.3). It follows that

$$\Pr(\mu_1 - \mu_2 \le 0 | \boldsymbol{x}, \boldsymbol{y}) = T_{m+n-2}(-u),$$

where T_{m+n-2} is the c.d.f. of the t distribution with $m+n-2$ degrees of freedom. Hence, the posterior probability that H_0 is true is less than $w_1/(w_0+w_1)$ if and only if

$$T_{m+n-2}(-u) < \frac{w_1}{w_0 + w_1}.$$

This, in turn is true if and only if

$$-u < T^{-1}_{m+n-2}\left(\frac{w_1}{w_0+w_1}\right).$$

This is true if and only if

$$u > T^{-1}_{m+n-2}\left(1 - \frac{w_1}{w_0+w_1}\right). \tag{9.8.18}$$

If $\alpha_0 = w_1/(w_0+w_1)$, then the Bayes test procedure that rejects H_0 when Eq. (9.8.18) occurs is the same as the level α_0 two-sample t test derived in Sec. 9.6. Put another way, the one-sided level α_0 two-sample t test rejects the null hypothesis H_0 if and only if the posterior probability that H_0 is true (based on the improper prior) is at most α_0. It follows from Exercise 7 that the posterior probability of the null hypothesis being true must equal the p-value in this case.

Example 9.8.6

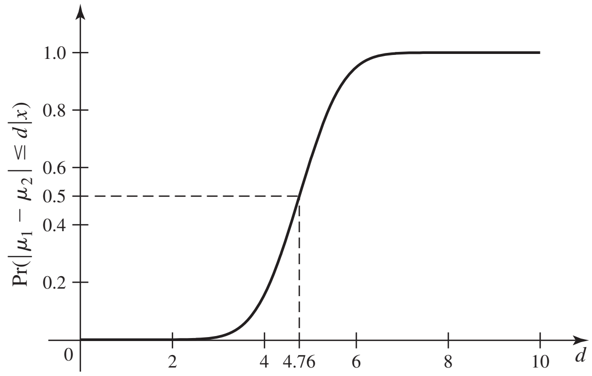
Roman Pottery in Britain. In Example 9.6.3, we observed 14 samples of Roman pottery from Llanederyn in Great Britain and another five samples from Ashley Rails, and we were interested in whether the mean aluminum oxide percentage in Llanederyn μ_1 was larger than that in Ashley Rails μ_2. We tested $H_0: \mu_1 \ge \mu_2$ against $H_1: \mu_1 < \mu_2$ and found that the p-value was 4×10^{-6}. If we had used an improper prior for the parameters, then $\Pr(\mu_1 \ge \mu_2 | \boldsymbol{x}) = 4 \times 10^{-6}$. ◀

Two-Sided Alternatives with Unknown Variance To test the hypothesis that the mean μ of a normal distribution is close to μ_0, we could specify a specific value d and test

$$\begin{aligned} H_0&: \quad |\mu - \mu_0| \le d, \\ H_1&: \quad |\mu - \mu_0| > d. \end{aligned}$$

If we do not feel comfortable selecting a single value of d to represent "close," we could compute $\Pr(|\mu - \mu_0| \le d | \boldsymbol{x})$ for all d and draw a plot as we did in Example 9.8.4.

Figure 9.18 Plot of $\Pr(|\mu_1 - \mu_2| \le d|\boldsymbol{x})$ against d. The dotted lines indicate that the median of the posterior distribution of $|\mu_1 - \mu_2|$ is 4.76.



The case of testing that two means are close together can be dealt with in the same way.

Example 9.8.7

Roman Pottery in Britain. In Example 9.8.6, we tested one-sided hypotheses about the difference in aluminum oxide contents in samples of pottery from two sites in Great Britain. Unless we are specifically looking for a difference in a particular direction, it might make more sense to test hypotheses of the form

$$\begin{aligned} H_0&: \quad |\mu_1 - \mu_2| \le d, \\ H_1&: \quad |\mu_1 - \mu_2| > d, \end{aligned} \tag{9.8.19}$$

where d is some critical difference that is worth detecting. As we did in Example 9.8.4, we can draw a plot that allows us to test all hypotheses of the form (9.8.19) simultaneously. We just plot $\Pr(|\mu_1 - \mu_2| \le d|\boldsymbol{x})$ against d. The posterior distribution of $\mu_1 - \mu_2$ was found in Eq. (9.8.17), using the improper prior. In this case, the following random variable has the t distribution with 17 degrees of freedom:

$$\begin{aligned} &(m+n-2)^{1/2} \frac{\mu_1 - \mu_2 - (\overline{x}_m - \overline{y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2} (s_x^2 + s_y^2)^{1/2}} \\ &= 17^{1/2} \frac{\mu_1 - \mu_2 - (12.56 - 17.32)}{\left(\frac{1}{14} + \frac{1}{5}\right)^{1/2} (24.65 + 11.01)^{1/2}} = 1.33(\mu_1 - \mu_2 + 4.76), \end{aligned}$$

where the data summaries come from Example 9.6.3. It follows that

$$\begin{aligned} &\Pr(|\mu_1 - \mu_2| \le d|\boldsymbol{x}) \\ &\quad = \Pr(1.33(-d + 4.76) \le 1.33(\mu_1 - \mu_2 + 4.76) \le 1.33(d + 4.76)|\boldsymbol{x}) \\ &\quad = T_{17}(1.33(d + 4.76)) - T_{17}(1.33(-d + 4.76)), \end{aligned}$$

where T_{17} is the c.d.f. of the t distribution with 17 degrees of freedom. Figure 9.18 is the plot of this posterior probability against d. ◄

Comparing the Variances of Two Normal Distributions

In order to test hypotheses concerning the variances of two normal distributions, we can make use of the posterior distribution of the ratio of the two variances. Suppose that $X_1, \ldots, X_m$ is a random sample from the normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, \ldots, Y_n$ is a random sample from the normal distribution with mean μ_2 and variance σ_2^2. If we model the X data and associated parameters as independent of the Y data and associated parameters, then we can perform two separate analyses just like the one in Sec. 8.6. In particular, we let $\tau_i = 1/\sigma_i^2$ for $i = 1, 2$, and the joint posterior distribution will have (μ_1, τ_1) independent of (μ_2, τ_2) and each pair will have a normal-gamma distribution just as in Sec. 8.6. For convenience, we shall only do the remaining calculations using improper priors. With improper priors, the posterior distribution of τ_1 is the gamma distribution with parameters $(m-1)/2$ and $s_x^2/2$, where s_x^2 is defined in Theorem 9.8.2. We also showed in Sec. 8.6 (using Exercise 1 in Sec. 5.7) that $\tau_1 s_x^2$ has the χ^2 distribution with $m-1$ degrees of freedom. Similarly, $\tau_2 s_y^2$ has the χ^2 distribution with $n-1$ degrees of freedom. Since $\tau_1 s_x^2/(m-1)$ and $\tau_2 s_y^2/(n-1)$ are independent, their ratio has the F distribution with $m-1$ and $n-1$ degrees of freedom. That is, the posterior distribution of

$$\frac{\tau_1 s_x^2/(m-1)}{\tau_2 s_y^2/(n-1)} = \frac{s_x^2/[(m-1)\sigma_1^2]}{s_y^2/[(n-1)\sigma_2^2]} \tag{9.8.20}$$

is the F distribution with $m-1$ and $n-1$ degrees of freedom. Notice that the expression on the right side of Eq. (9.8.20) is the same as the random variable V^* in Eq. (9.7.5). This is another case in which the sampling distribution of a random variable is the same as its posterior distribution. It will then follow that level α_0 tests of one-sided hypotheses about σ_1^2/σ_2^2 based on the sampling distribution of V^* will be the same as Bayes tests of the form (9.8.7) so long as $\alpha_0 = w_1/(w_0 + w_1)$. The reader can prove this in Exercise 9.

Summary

From a Bayesian perspective, one chooses a test procedure by minimizing the posterior expected loss. When the loss has the simple form of (9.8.6), then the Bayes test procedure is to reject H_0 when its posterior probability is at most $w_1/(w_0 + w_1)$. In many one-sided cases, with improper priors, this procedure turns out to be the same as the most commonly used level $\alpha_0 = w_1/(w_0 + w_1)$ test. In two-sided cases, as an alternative to testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$, one can draw a plot of $\Pr(|\theta - \theta_0| \le d|\boldsymbol{x})$ against d. One then needs to decide which values of d count as meaningful differences.

Exercises

1. Suppose that a certain industrial process can be either in control or out of control, and that at any specified time the prior probability that it will be in control is 0.9, and the prior probability that it will be out of control is 0.1. A single observation X of the output of the process is to be taken, and it must be decided immediately whether the process is in control or out of control. If the process is in control, then X will have the normal distribution with mean 50 and variance 1.If the process is out of control, then X will have the normal distribution with mean 52 and variance 1.

If it is decided that the process is out of control when in fact it is in control, then the loss from unnecessarily stopping the process will be \$1000. If it is decided that the process is in control when in fact it is out of control, then the loss from continuing the process will be \$18,000. If a correct decision is made, then the loss will be 0. It is desired to find a test procedure for which the expected loss will be a minimum. For what values of X should it be decided that the process is out of control?

2. A single observation X is to be taken from a continuous distribution for which the p.d.f. is either f_0 or f_1, where

$$f_0(x) = \begin{cases} 1 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_1(x) = \begin{cases} 4x^3 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

On the basis of the observation X, it must be decided whether f_0 or f_1 is the correct p.d.f. Suppose that the prior probability that f_0 is correct is 2/3 and the prior probability that f_1 is correct is 1/3. Suppose also that the loss from choosing the correct decision is 0, the loss from deciding that f_1 is correct when in fact f_0 is correct is 1 unit, and the loss from deciding that f_0 is correct when in fact f_1 is correct is 4 units. If the expected loss is to be minimized, for what values of X should it be decided that f_0 is correct?

3. Suppose that a failure in a certain electronic system can occur because of either a minor or a major defect. Suppose also that 80 percent of the failures are caused by minor defects, and 20 percent of the failures are caused by major defects. When a failure occurs, n independent soundings $X_1, \ldots, X_n$ are made on the system. If the failure was caused by a minor defect, these soundings form a random sample from the Poisson distribution with mean 3. If the failure was caused by a major defect, these soundings form a random sample from a Poisson distribution for which the mean is 7. The cost of deciding that the failure was caused by a major defect when it was actually caused by a minor defect is \$400. The cost of deciding that the failure was caused by a minor defect when it was actually caused by a major defect is \$2500. The cost of choosing a correct decision is 0. For a given set of observed values of $X_1, \ldots, X_n$, which decision minimizes the expected cost?

4. Suppose that the proportion p of defective items in a large manufactured lot is unknown, and it is desired to test the following simple hypotheses:

$$H_0: \quad p = 0.3,$$
$$H_1: \quad p = 0.4.$$

Suppose that the prior probability that $p = 0.3$ is 1/4, and the prior probability that $p = 0.4$ is 3/4; also suppose that the loss from choosing an incorrect decision is 1 unit, and the loss from choosing a correct decision is 0. Suppose that a random sample of n items is selected from the lot. Show that the Bayes test procedure is to reject H_0 if and only if the proportion of defective items in the sample is greater than

$$\frac{\log\left(\frac{7}{6}\right) + \frac{1}{n}\log\left(\frac{1}{3}\right)}{\log\left(\frac{14}{9}\right)}.$$

5. Suppose that we wish to test the hypotheses (9.8.1). Let the loss function have the form of (9.8.2).

a. Prove that the posterior probability of $\theta = \theta_0$ is $\xi_0 f_0(\boldsymbol{x})/[\xi_0 f_0(\boldsymbol{x}) + \xi_1 f_1(\boldsymbol{x})]$.

b. Prove that a test that minimizes $r(\delta)$ also minimizes the posterior expected value of the loss given $\boldsymbol{X} = \boldsymbol{x}$ for all $\boldsymbol{x}$.

c. Prove that the following test is one of the tests described in part (b): "reject H_0 if $\Pr(H_0 \text{ true}|\boldsymbol{x}) \le w_1/(w_0 + w_1)$."

6. Prove that the conclusion of Theorem 9.8.1 still holds when the loss function is given by

	d_0	d_1
$\theta \le \theta_0$	0	$w_0(\theta)$
$\theta > \theta_0$	$w_1(\theta)$	0

for arbitrary positive functions $w_0(\theta)$ and $w_1(\theta)$. *Hint:* Replicate the proof of Theorem 9.8.1, but replace the constants w_0 and w_1 by the functions above and keep them inside of the integrals instead of factoring them out.

7. Suppose that we have a situation in which the Bayes test that rejects H_0 when $\Pr(H_0 \text{ true }|\boldsymbol{x}) \le \alpha_0$ is the same as the level α_0 test of H_0 for all α_0. (Example 9.8.5 has this property, but so do many other situations.) Prove that the p-value equals the posterior probability that H_0 is true.

8. In this exercise you will prove Theorem 9.8.2.

a. Prove that the joint p.d.f. of the data given the parameters μ_1, μ_2, and τ can be written as a constant times

$$\tau^{(m+n)/2} \exp\Big(-0.5m\tau(\mu_1 - \overline{x}_m)^2 - 0.5n\tau(\mu_2 - \overline{y}_n)^2 - 0.5(s_x^2 + s_y^2)\tau\Big).$$

b. Multiply the prior p.d.f. times the p.d.f. in part (a). Bayes' theorem for random variables says that the result is proportional (as a function of the parameters) to the posterior p.d.f.

i. Show that the posterior p.d.f., as a function of μ_1 for fixed μ_2 and τ, is the p.d.f. of the normal distribution with mean $\overline{x}_m$ and variance $(m\tau)^{-1}$.

ii. Show that the posterior p.d.f., as a function of μ_2 for fixed μ_1 and τ, is the p.d.f. of the normal distribution with mean $\overline{y}_n$ and variance $(n\tau)^{-1}$.

iii. Show that, conditional on τ, μ_1 and μ_2 are independent with the two normal distributions found above.

iv. Show that the marginal posterior distribution of τ is the gamma distribution with parameters $(m + n - 2)/2$ and $(s_x^2 + s_y^2)/2$.

c. Show that the conditional distribution of

$$Z = \tau^{1/2} \frac{\mu_1 - \mu_2 - (\overline{x}_m - \overline{y}_n)}{\left(\dfrac{1}{m} + \dfrac{1}{n}\right)^{1/2}}$$

given τ is a standard normal distribution and hence Z is independent of τ.

d. Show that the distribution of $W = (s_x^2 + s_y^2)\tau$ is the gamma distribution with parameters $(m + n - 2)/2$ and $1/2$, which is the same as the χ^2 distribution with $m + n - 2$ degrees of freedom.

e. Prove that $Z/(W/(m + n - 2))^{1/2}$ has the t distribution with $m + n - 2$ degrees of freedom and that it equals the expression in Eq. (9.8.17).

9. Suppose that $X_1, \dots, X_m$ form a random sample from the normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, \dots, Y_n$ form a random sample from the normal distribution with mean μ_2 and variance σ_2^2. Suppose that we use the usual improper prior and that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \sigma_1^2 \le \sigma_2^2, \\ H_1&: \quad \sigma_1^2 > \sigma_2^2. \end{aligned}$$

a. Prove that the level α_0 F test is the same as the test in (9.8.7) when $\alpha_0 = w_1/(w_0 + w_1)$.

b. Prove that the p-value for the F test is the posterior probability that H_0 is true.

10. Consider again the situation in Example 9.6.2. Let μ_1 be the mean of log-rainfall from seeded clouds, and let μ_2 be the mean of log-rainfall from unseeded clouds. Use the improper prior for the parameters.

a. Find the posterior distribution of $\mu_1 - \mu_2$.

b. Draw a graph of the posterior probability that $|\mu_1 - \mu_2| \le d$ as a function of d.

11. Let θ be a general parameter taking values in a parameter space Ω. Let $\Omega' \bigcup \Omega'' = \Omega$ be a partition of Ω into two disjoint sets Ω' and Ω''. We want to choose between two decisions: d' says that $\theta \in \Omega'$, and d'' says that $\theta \in \Omega''$. We have the following loss function:

	d'	d''
If $\theta \in \Omega'$	0	w'
If $\theta \in \Omega''$	w''	0

We have two choices for expressing this decision problem as a hypothesis-testing problem. One choice would be to define $H_0: \theta \in \Omega'$ and $H_1: \theta \in \Omega''$. The other choice would be to define $H_0: \theta \in \Omega''$ and $H_1: \theta \in \Omega'$. In this problem, we show that the Bayes test makes the same decision regardless of which hypothesis we call the null and which we call the alternative.

a. For each choice, say how we would define each of the following in order to make this problem fit the hypothesis-testing framework described in this section: w_0, w_1, d_0, d_1, Ω_0, and Ω_1.

b. Now suppose that we can observe data $\boldsymbol{X} = \boldsymbol{x}$ and compute the posterior distribution of θ, $\xi(\theta|\boldsymbol{x})$. Show that, for each of the two setups constructed in the previous part, the Bayes test chooses the same decision d' or d''. That is, observing $\boldsymbol{x}$ leads to choosing d' in the first setup if and only if observing $\boldsymbol{x}$ leads to choosing d' in the second setup. Similarly, observing $\boldsymbol{x}$ leads to choosing d'' in the first setup if and only if observing $\boldsymbol{x}$ leads to choosing d'' in the second setup.

⋆ 9.9 Foundational Issues

We discuss the relationship between significance level and sample size. We also distinguish between results that are significant in the statistical sense and those that are significant in a practical sense.

The Relationship between Level of Significance and Sample Size

In many statistical applications, it has become standard practice for an experimenter to specify a level of significance α_0, and then to find a test procedure with a large power function on the alternative hypothesis among all procedures whose size $\alpha(\delta) \le \alpha_0$. Alternatively, the experimenter will compute a p-value and report whether or not it was less than α_0. For the case of testing simple null and alternative hypotheses, the Nayman-Pearson lemma explicitly describes how to construct such a procedure. Furthermore, it has become traditional in many applications to choose the level of significance α_0 to be 0.10, 0.05, or 0.01. The selected level depends on how serious the consequences of an error of type I are judged to be. The value of α_0 most commonly used is 0.05. If the consequences of an error of type I are judged to be relatively mild in a particular problem, the experimenter may choose α_0 to be 0.10. On the other hand, if these consequences are judged to be especially serious, the experimenter may choose α_0 to be 0.01.

Because these values of α_0 have become established in statistical practice, the choice of $\alpha_0 = 0.01$ is sometimes made by an experimenter who wishes to use a cautious test procedure, or one that will not reject H_0 unless the sample data provide strong evidence that H_0 is not true. We shall now show, however, that when the sample size n is large, the choice of $\alpha_0 = 0.01$ can actually lead to a test procedure that will reject H_0 for certain samples that, in fact, provide stronger evidence for H_0 than they do for H_1.

To illustrate this property, suppose, as in Example 9.2.5, that a random sample is taken from the normal distribution with unknown mean θ and known variance 1, and that the hypotheses to be tested are

$$\begin{aligned} H_0&: \quad \theta = 0, \\ H_1&: \quad \theta = 1. \end{aligned}$$

It follows from the discussion in Example 9.2.5 that, among all test procedures for which $\alpha(\delta) \le 0.01$, the probability of type II error $\beta(\delta)$ will be a minimum for the procedure δ^* that rejects H_0 when $\overline{X}_n \ge k'$, where k' is chosen so that $\Pr(\overline{X}_n \ge k'|\theta = 0) = 0.01$. When $\theta = 0$, the random variable $\overline{X}_n$ has the normal distribution with mean 0 and variance $1/n$. Therefore, it can be found from a table of the standard normal distribution that $k' = 2.326n^{-1/2}$.

Furthermore, it follows from Eq. (9.2.12) that this test procedure δ^* is equivalent to rejecting H_0 when $f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) \ge k$, where $k = \exp(2.326n^{1/2} - 0.5n)$. The probability of an error of type I will be $\alpha(\delta^*) = 0.01$. Also, by an argument similar to the one leading to Eq. (9.2.15), the probability of an error of type II will be $\beta(\delta^*) = \Phi(2.326 - n^{1/2})$, where Φ denotes the c.d.f. of the standard normal distribution. For $n = 1, 25$, and 100, the values of $\beta(\delta^*)$ and k are as follows:

n	$\alpha(\delta^*)$	$\beta(\delta^*)$	k
1	0.01	0.91	6.21
25	0.01	0.0038	0.42
100	0.01	8×10^{-15}	2.5×10^{-12}

It can be seen from this tabulation that when $n = 1$, the null hypothesis H_0 will be rejected only if the likelihood ratio $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ exceeds the value $k = 6.21$. In other words, H_0 will not be rejected unless the observed values $x_1, \ldots, x_n$ in the sample are at least 6.21 times as likely under H_1 as they are under H_0. In this case, the procedure δ^* therefore satisfies the experimenter's desire to use a test that is cautious about rejecting H_0.

If $n = 100$, however, the procedure δ^* will reject H_0 whenever the likelihood ratio exceeds the value $k = 2.5 \times 10^{-12}$. Therefore, H_0 will be rejected for certain observed values $x_1, \ldots, x_n$ that are actually millions of times more likely under H_0 as they are under H_1. The reason for this result is that the value of $\beta(\delta^*)$ that can be achieved when $n = 100$, which is 8×10^{-15}, is extremely small relative to the specified value $\alpha_0 = 0.01$. Hence, the procedure δ^* actually turns out to be much more cautious about an error of type II than it is about an error of type I. We can see from this discussion that a value of α_0 that is an appropriate choice for a small value of n might be unnecessarily large for a large value of n. Hence, it would be sensible to let the level of significance α_0 decrease as the sample size increases.

Suppose now that the experimenter regards an error of type I to be much more serious than an error of type II, and she therefore desires to use a test procedure for which the value of the linear combination $100\alpha(\delta) + \beta(\delta)$ will be a minimum. Then it follows from Theorem 9.2.1 that she should reject H_0 if and only if the likelihood ratio exceeds the value $k = 100$, regardless of the sample size n. In other words, the procedure that minimizes the value of $100\alpha(\delta) + \beta(\delta)$ will not reject H_0 unless the observed values $x_1, \ldots, x_n$ are at least 100 times as likely under H_1 as they are under H_0.

From this discussion, it seems more reasonable for the experimenter to take the values of both $\alpha(\delta)$ and $\beta(\delta)$ into account when choosing a test procedure, rather than to fix a value of $\alpha(\delta)$ and minimize $\beta(\delta)$. For example, one could minimize the value of a linear combination of the form $a\alpha(\delta) + b\beta(\delta)$. In Sec. 9.8, we saw how the Bayesian point of view also leads to the conclusion that one should try to minimize a linear combination of this form. Lehmann (1958) suggested choosing a number k and requiring that $\beta(\delta) = k\alpha(\delta)$. Both the Bayesian method and Lehmann's method have the advantage of forcing the probabilities of both type I and type II errors to decrease as one obtains more data. Similar problems with fixing the significance level of a test arise when hypotheses are composite, as we illustrate later in this section.

Statistically Significant Results

When the observed data lead to rejecting a null hypothesis H_0 at level α_0, it is often said that one has obtained a result that is *statistically significant* at level α_0. When this occurs, it does not mean that the experimenter should behave as if H_0 is false. Similarly, if the data do not lead to rejecting H_0, the result is not statistically significant at level α_0, but the experimenter should not necessarily become convinced that H_0 is true. Indeed, qualifying "significant" with the term "statistically" is a warning that a statistically significant result might be different than a practically significant result. Consider, once again, Example 9.5.10 on page 582, in which the hypotheses to be tested are

$$\begin{aligned} H_0&: \quad \mu = 5.2, \\ H_1&: \quad \mu \neq 5.2. \end{aligned}$$

It is extremely important for the experimenter to distinguish a statistically significant result from any claim that the parameter μ is significantly different from the hypothesized value 5.2. Even if the data suggest that μ is not equal to 5.2, this does not necessarily provide any evidence that the actual value of μ is *significantly* different from 5.2. For a given set of data, the tail area corresponding to the observed value of the test statistic U might be very small, and yet the data might suggest that the actual value of μ is so close to 5.2 that, for practical purposes, the experimenter would not regard μ as being significantly different from 5.2.

The situation just described can arise when the statistic U is based on a very large random sample. Suppose, for instance, that in Example 9.5.10 the lengths of 20,000 fibers in a random sample are measured, rather than the lengths of only 15 fibers. For a given level of significance, say, $\alpha_0 = 0.05$, let $\pi(\mu, \sigma^2|\delta)$ denote the power function of the t test based on these 20,000 observations. Then $\pi(5.2, \sigma^2|\delta) = 0.05$ for every value of $\sigma^2 > 0$. However, because of the very large number of observations on which the test is based, the power $\pi(\mu, \sigma^2|\delta)$ will be very close to 1 for each value of μ that differs only slightly from 5.2 and for a moderate value of σ^2. In other words, even if the value of μ differs only slightly from 5.2, the probability is close to 1 that one would obtain a statistically significant result. For example, with $n = 20{,}000$, the power of the level 0.05 test when $|\mu - 5.2| = 0.03\sigma$ is 0.99.

As explained in Sec. 9.4, it is inconceivable that the mean length μ of all the fibers in the entire population will be exactly 5.2. However, μ may be very close to 5.2, and when it is, the experimenter will not want to reject the null hypothesis H_0. Nevertheless, it is very likely that the t test based on the sample of 20,000 fibers will lead to a statistically significant result. Therefore, when an experimenter analyzes a powerful test based on a very large sample, he must exercise caution in interpreting the actual significance of a "statistically significant" result. He knows in advance that there is a high probability of rejecting H_0 even when the true value of μ differs only slightly from the value 5.2 specified under H_0.

One way to handle this situation, as discussed earlier in this section, is to recognize that a level of significance much smaller than the traditional value of 0.05 or 0.01 is appropriate for a problem with a large sample size. Another way is to replace the single value of μ in the null hypothesis by an interval, as we did on pages 571 and 610. A third way is to regard the statistical problem as one of estimation rather than one of testing hypotheses.

When a large random sample is available, the sample mean and the sample variance will be excellent estimators of the parameters μ and σ^2. Before the experimenter chooses any decision involving the unknown values of μ and σ^2, she should calculate and consider the values of these estimators as well as the value of the statistic U.

Summary

When we reject a null hypothesis, we say that we have obtained a statistically significant result. The power function of a level α_0 test becomes very large, even for parameter values close to the null hypothesis, as the size of the sample increases. For the case of simple hypotheses, the probability of type II error can become very small while the probability of type I error stays as large as α_0. One way to avoid this is to let the level of significance decrease as the sample size increases. If one rejects a null hypothesis at a particular level of significance α_0, one must be careful to check whether the data actually suggest any deviation of practical importance from the null hypothesis.

Exercises

1. Suppose that a single observation X is taken from the normal distribution with unknown mean μ and known variance is 1. Suppose that it is known that the value of μ must be -5, 0, or 5, and it is desired to test the following hypotheses at the level of significance 0.05:

$$H_0:\ \mu = 0,$$
$$H_1:\ \mu = -5 \text{ or } \mu = 5.$$

Suppose also that the test procedure to be used specifies rejecting H_0 when $|X| > c$, where the constant c is chosen so that $\Pr(|X| > c|\mu = 0) = 0.05$.

a. Find the value of c, and show that if $X = 2$, then H_0 will be rejected.

b. Show that if $X = 2$, then the value of the likelihood function at $\mu = 0$ is 12.2 times as large as its value at $\mu = 5$ and is 5.9×10^9 times as large as its value at $\mu = -5$.

2. Suppose that a random sample of 10,000 observations is taken from the normal distribution with unknown mean μ and known variance is 1, and it is desired to test the following hypotheses at the level of significance 0.05:

$$H_0:\ \mu = 0,$$
$$H_1:\ \mu \neq 0.$$

Suppose also that the test procedure specifies rejecting H_0 when $|\overline{X}_n| \geq c$, where the constant c is chosen so that $\Pr(|\overline{X}_n| \geq c|\mu = 0) = 0.05$. Find the probability that the test will reject H_0 if **(a)** the actual value of μ is 0.01, and **(b)** the actual value of μ is 0.02.

3. Consider again the conditions of Exercise 2, but suppose now that it is desired to test the following hypotheses:

$$H_0:\quad \mu \le 0,$$
$$H_1:\quad \mu > 0.$$

Suppose also that in the random sample of 10,000 observations, the sample mean $\overline{X}_n$ is 0.03. At what level of significance is this result just significant?

4. Suppose that $X_1, \dots, X_n$ comprise a random sample from the normal distribution with unknown mean θ and known variance 1. Suppose that it is desired to test the same hypotheses as in Exercise 3. This time, however, the test procedure δ will be chosen so as to minimize $19\pi(0|\delta) + 1 - \pi(0.5|\delta)$.

a. Find the value c_n so that the test procedure δ rejects H_0 if $\overline{X}_n \ge c_n$ for each value $n = 1$, $n = 100$, and $n = 10{,}000$.

b. For each value of n in part (a), find the size of the test procedure δ.

5. Suppose that $X_1, \dots, X_n$ comprise a random sample from the normal distribution with unknown mean θ and variance 1. Suppose that it is desired to test the same hypotheses as in Exercise 3. This time, however, the test procedure δ will be chosen so that $19\pi(0|\delta) = 1 - \pi(0.5|\delta)$.

a. Find the value c_n so that the test procedure δ rejects H_0 if $\overline{X}_n \ge c_n$ for each value $n = 1$, $n = 100$, and $n = 10{,}000$.

b. For each value of n in part (a), find the size of the test procedure δ.

9.10 Supplementary Exercises

1. I will flip a coin three times and let X stand for the number of times that the coin comes up heads. Let θ stand for the probability that the coin comes up heads on a single flip, and assume that the flips are independent given θ. I wish to test the null hypothesis $H_0: \theta = 1/2$ against the alternative hypothesis $H_1: \theta = 3/4$. Find the test δ that minimizes $\alpha(\delta) + \beta(\delta)$, the sum of the type I and type II error probabilities, and find the two error probabilities for the test.

2. Suppose that a sequence of Bernoulli trials is to be carried out with an unknown probability θ of success on each trial, and the following hypotheses are to be tested:

$$H_0:\quad \theta = 0.1,$$
$$H_1:\quad \theta = 0.2.$$

Let X denote the number of trials required to obtain a success, and suppose that H_0 is to be rejected if $X \le 5$. Determine the probabilities of errors of type I and type II.

3. Consider again the conditions of Exercise 2. Suppose that the losses from errors of type I and type II are equal, and the prior probabilities that H_0 and H_1 are true are equal. Determine the Bayes test procedure based on the observation X.

4. Suppose that a single observation X is to be drawn from the following p.d.f.:

$$f(x|\theta) = \begin{cases} 2(1-\theta)x + \theta & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

where the value of θ is unknown $(0 \le \theta \le 2)$. Suppose also that the following hypotheses are to be tested:

$$H_0:\quad \theta = 2,$$
$$H_1:\quad \theta = 0.$$

Determine the test procedure δ for which $\alpha(\delta) + 2\beta(\delta)$ is a minimum, and calculate this minimum value.

5. Consider again the conditions of Exercise 4, and suppose that $\alpha(\delta)$ is required to be a given value α_0 $(0 < \alpha_0 < 1)$. Determine the test procedure δ for which $\beta(\delta)$ will be a minimum, and calculate this minimum value.

6. Consider again the conditions of Exercise 4, but suppose now that the following hypotheses are to be tested:

$$H_0:\quad \theta \ge 1,$$
$$H_1:\quad \theta < 1.$$

a. Determine the power function of the test δ that specifies rejecting H_0 if $X \ge 0.9$.

b. What is the size of the test δ?

7. Consider again the conditions of Exercise 4. Show that the p.d.f. $f(x|\theta)$ has a monotone likelihood ratio in the statistic $r(X) = -X$, and determine a UMP test of the following hypotheses at the level of significance $\alpha_0 = 0.05$:

$$H_0:\quad \theta \le \tfrac{1}{2},$$
$$H_1:\quad \theta > \tfrac{1}{2}.$$

8. Suppose that a box contains a large number of chips of three different colors, red, brown, and blue, and it is desired to test the null hypothesis H_0 that chips of the three colors are present in equal proportions against the alternative hypothesis H_1 that they are not present in equal proportions. Suppose that three chips are to be drawn at random from the box, and H_0 is to be rejected if and only if at least two of the chips have the same color.

a. Determine the size of the test.

b. Determine the power of the test if 1/7 of the chips are red, 2/7 are brown, and 4/7 are blue.

9. Suppose that a single observation X is to be drawn from an unknown distribution P, and that the following simple hypotheses are to be tested:

H_0: P is the uniform distribution on the interval [0, 1],

H_1: P is the standard normal distribution.

Determine the most powerful test of size 0.01, and calculate the power of the test when H_1 is true.

10. Suppose that the 12 observations $X_1, \dots, X_{12}$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Describe how to carry out a t test of the following hypotheses at the level of significance $\alpha_0 = 0.005$:

$$\begin{aligned} H_0&: \quad \mu \geq 3,\\ H_1&: \quad \mu < 3. \end{aligned}$$

11. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean θ and known variance 1, and it is desired to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \theta \leq 0,\\ H_1&: \quad \theta > 0. \end{aligned}$$

Suppose also that it is decided to use a UMP test for which the power is 0.95 when $\theta = 1$. Determine the size of this test if $n = 16$.

12. Suppose that eight observations $X_1, \dots, X_8$ are drawn at random from a distribution with the following p.d.f.:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1,\\ 0 & \text{otherwise.} \end{cases}$$

Suppose also that the value of θ is unknown ($\theta > 0$), and it is desired to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \theta \leq 1,\\ H_1&: \quad \theta > 1. \end{aligned}$$

Show that the UMP test at the level of significance $\alpha_0 = 0.05$ specifies rejecting H_0 if $\sum_{i=1}^{8} \log X_i \geq -3.981$.

13. Suppose that $X_1, \dots, X_n$ form a random sample from the χ^2 distribution with unknown degrees of freedom θ ($\theta = 1, 2, \dots$), and it is desired to test the following hypotheses at a given level of significance α_0 ($0 < \alpha_0 < 1$):

$$\begin{aligned} H_0&: \quad \theta \leq 8,\\ H_1&: \quad \theta \geq 9. \end{aligned}$$

Show that there exists a UMP test, and the test specifies rejecting H_0 if $\sum_{i=1}^{n} \log X_i \geq k$ for some appropriate constant k.

14. Suppose that $X_1, \dots, X_{10}$ form a random sample from a normal distribution for which both the mean and the variance are unknown. Construct a statistic that does not depend on any unknown parameters and has the F distribution with three and five degrees of freedom.

15. Suppose that $X_1, \dots, X_m$ form a random sample from the normal distribution with unknown mean μ_1 and unknown variance σ_1^2, and that $Y_1, \dots, Y_n$ form an independent random sample from the normal distribution with unknown mean μ_2 and unknown variance σ_2^2. Suppose also that it is desired to test the following hypotheses with the usual F test at the level of significance $\alpha_0 = 0.05$:

$$\begin{aligned} H_0&: \quad \sigma_1^2 \leq \sigma_2^2,\\ H_1&: \quad \sigma_1^2 > \sigma_2^2. \end{aligned}$$

Assuming that $m = 16$ and $n = 21$, show that the power of the test when $\sigma_1^2 = 2\sigma_2^2$ is given by $\Pr(V^* \geq 1.1)$, where V^* is a random variable having the F distribution with 15 and 20 degrees of freedom.

16. Suppose that the nine observations $X_1, \dots, X_9$ form a random sample from the normal distribution with unknown mean μ_1 and unknown variance σ^2, and the nine observations $Y_1, \dots, Y_9$ form an independent random sample from the normal distribution with unknown mean μ_2 and the same unknown variance σ^2. Let S_X^2 and S_Y^2 be as defined in Eq. (9.6.2) (with $m = n = 9$), and let

$$T = \max\left\{ \frac{S_X^2}{S_Y^2}, \frac{S_Y^2}{S_X^2} \right\}.$$

Determine the value of the constant c such that $\Pr(T > c) = 0.05$.

17. An unethical experimenter desires to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \theta = \theta_0,\\ H_1&: \quad \theta \neq \theta_0. \end{aligned}$$

She draws a random sample $X_1, \dots, X_n$ from a distribution with the p.d.f. $f(x|\theta)$, and carries out a test of size α. If this test does not reject H_0, she discards the sample, draws a new independent random sample of n observations, and repeats the test based on the new sample. She continues drawing new independent samples in this way until she obtains a sample for which H_0 is rejected.

a. What is the overall size of this testing procedure?

b. If H_0 is true, what is the expected number of samples that the experimenter will have to draw until she rejects H_0?

18. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown precision τ, and the following hypotheses are to be tested:

$$H_0: \quad \mu \le 3,$$
$$H_1: \quad \mu > 3.$$

Suppose that the prior joint distribution of μ and τ is the normal-gamma distribution, as described in Theorem 8.6.1, with $\mu_0 = 3$, $\lambda_0 = 1$, $\alpha_0 = 1$, and $\beta_0 = 1$. Suppose finally that $n = 17$, and it is found from the observed values in the sample that $\overline{X}_n = 3.2$ and $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 17$. Determine both the prior probability and the posterior probability that H_0 is true.

19. Consider a problem of testing hypotheses in which the following hypotheses about an arbitrary parameter θ are to be tested:

$$H_0: \quad \theta \in \Omega_0,$$
$$H_1: \quad \theta \in \Omega_1.$$

Suppose that δ is a test procedure of size α $(0 < \alpha < 1)$ based on some vector of observations $\boldsymbol{X}$, and let $\pi(\theta|\delta)$ denote the power function of δ. Show that if δ is unbiased, then $\pi(\theta|\delta) \ge \alpha$ at every point $\theta \in \Omega_1$.

20. Consider again the conditions of Exercise 19. Suppose now that we have a two-dimensional vector $\boldsymbol{\theta} = (\theta_1, \theta_2)$, where θ_1 and θ_2 are real-valued parameters. Suppose also that A is a particular circle in the $\theta_1\theta_2$-plane, and that the hypotheses to be tested are as follows:

$$H_0: \quad \theta \in A,$$
$$H_1: \quad \theta \notin A.$$

Show that if the test procedure δ is unbiased and of size α, and if its power function $\pi(\theta|\delta)$ is a continuous function of θ, then it must be true that $\pi(\theta|\delta) = \alpha$ at each point θ on the boundary of the circle A.

21. Consider again the conditions of Exercise 19. Suppose now that θ is a real-valued parameter, and the following hypotheses are to be tested:

$$H_0: \quad \theta = \theta_0,$$
$$H_1: \quad \theta \ne \theta_0.$$

Assume that θ_0 is an interior point of the parameter space Ω. Show that if the test procedure δ is unbiased and if its power function $\pi(\theta|\delta)$ is a differentiable function of θ, then $\pi'(\theta_0|\delta) = 0$, where $\pi'(\theta_0|\delta)$ denotes the derivative of $\pi(\theta|\delta)$ evaluated at the point $\theta = \theta_0$.

22. Suppose that the differential brightness θ of a certain star has an unknown value, and it is desired to test the following simple hypotheses:

$$H_0: \quad \theta = 0,$$
$$H_1: \quad \theta = 10.$$

The statistician knows that when he goes to the observatory at midnight to measure θ, there is probability 1/2 that the meteorological conditions will be good, and he will be able to obtain a measurement X having the normal distribution with mean θ and variance 1. He also knows that there is probability 1/2 that the meteorological conditions will be poor, and he will obtain a measurement Y having the normal distribution with mean θ and variance 100. The statistician also learns whether the meteorological conditions were good or poor.

- **a.** Construct the most powerful test that has conditional size $\alpha = 0.05$, given good meteorological conditions, and one that has conditional size $\alpha = 0.05$, given poor meteorological conditions.
- **b.** Construct the most powerful test that has conditional size $\alpha = 2.0 \times 10^{-7}$, given good meteorological conditions, and one that has conditional size $\alpha = 0.0999998$, given poor meteorological conditions. (You will need a computer program to do this.)
- **c.** Show that the overall size of both the test found in part (a) and the test found in part (b) is 0.05, and determine the power of each of these two tests.

23. Consider again the situation described in Exercise 22. This time, assume that there is a loss function of the form (9.8.6). Also, assume that the prior probability of $\theta = 0$ is ξ_0 and the prior probability of $\theta = 10$ is ξ_1.

- **a.** Find the formula for the Bayes test for general loss function of the form (9.8.6).
- **b.** Prove that the test in part (a) of Exercise 22 is not a special case of the Bayes test found in part (a) of the present exercise.
- **c.** Prove that the test in part (b) of Exercise 22 is (up to rounding error) a special case of the Bayes test found in part (a) of the present exercise.

24. Let $X_1, \ldots, X_n$ be i.i.d. with the Poisson distribution having mean θ. Let $Y = \sum_{i=1}^{n} X_i$.

- **a.** Suppose that we wish to test the hypotheses $H_0 : \theta \ge 1$ versus $H_1 : \theta < 1$. Show that the test "reject H_0 if $Y = 0$" is uniformly most powerful level α_0 for some number α_0. Also find α_0.
- **b.** Find the power function of the test from part (a).

25. Consider a family of distributions with parameter θ and monotone likelihood ratio in a statistic T. We learned how to find a uniformly most powerful level α_0 test δ_c of the null hypothesis $H_{0,c} : \theta \le c$ versus $H_{1,c} : \theta > c$ for every c. We also know that these tests are equivalent to a coefficient $1 - \alpha_0$ confidence interval, where the confidence interval contains c if and only if δ_c does not reject $H_{0,c}$. The confidence interval is called *uniformly most accurate coefficient* $1 - \alpha_0$. Based on the equivalence of the tests and the confidence interval, figure out what the definition of "uniformly most accurate coefficient $1 - \alpha_0$" must be. Write the definition in terms of the conditional probability that the interval covers θ_1 given that $\theta = \theta_2$ for various pairs of values θ_1 and θ_2.