8

Tests of Hypotheses

In Chapter 7 we considered problems of estimation; in this chapter we will study what are generally called tests of hypotheses. A common dictionary definition of the word hypothesis states a hypothesis is "an unproved theory, proposition, supposition." For our purposes we will be concerned with hypotheses about probability laws; by observing values of the random variable whose probability law is affected, we gather evidence regarding the truth of the hypothesis.

For example, in Example 7.4.2 we assumed the time to failure for a component in "standard use" to be an exponential random variable with parameter λ, and assumed if this same component were tested in a more severe environment, its time to failure is again exponential, but with parameter 4λ. This situation could be cast as defining a hypothesis: Probability law for time to failure, standard use, is exponential, parameter λ_1. Probability law for time to failure, severe environment, is exponential, parameter λ_2. It could be the case that $\lambda_2 = 4\lambda_1$ (this is the hypothesis). If we were to observe n independent times to failure, $X_1, X_2, \ldots, X_n$, for the standard use and independent times to failure, $Y_1, Y_2, \ldots, Y_m$, in the severe environment, the observed values of the random variables should contain information we could use to decide whether or not the hypothesis ($\lambda_2 = 4\lambda_1$) appears to be true. We will examine some of the classical methodology for accomplishing this in the present chapter.

By defining a "test of a hypothesis," we will simply mean we have specified a rule that, for any possible collection of observed values for the random variables, tells us whether or not to accept the hypothesis. There are, of course, any number of possible rules (tests) that could be employed for any given problem. We will seek to find the rule that may be best in some sense. For example, in the life-testing case, recall that $\bar{X}$ estimates $1/\lambda_1$ (the mean time to failure, standard use) whereas $\bar{Y}$ estimates $1/\lambda_2$ (mean time to failure in the severe environment). If in actual fact it is true that $\lambda_2 = 4\lambda_1$, then we would expect to find $\bar{X}/\bar{Y}$ equal to approximately $(1/\lambda_1)/(1/4\lambda_1) = 4$. But since both $\bar{X}$ and $\bar{Y}$ are (continuous) random variables, and both would vary from one collection of possible observed values to another, we realize that we will not find $\bar{x}/\bar{y} = 4$ for all possible samples, even if $\lambda_2 = 4\lambda_1$. Thus the thought that might immediately occur, that we accept the hypothesis $\lambda_2 = 4\lambda_1$, only if we find $\bar{x}/\bar{y} = 4$, really is not totally practical. The variability in both $\bar{X}$ and $\bar{Y}$ makes a range of values for $\bar{x}/\bar{y}$ possible, when $\lambda_2 = 4\lambda_1$; thus we probably should settle on some rule (test) like "Accept the hypothesis $\lambda_2 = 4\lambda_1$ if $\bar{x}/\bar{y}$ is close to 4," where it remains to be seen what might be meant by the phrase "close to 4." As we will see, this is in fact the best type of rule from several points of view.

The observed values of random samples contain information about the probability law, but they do not contain "perfect" information, because the same set of observed values could come from many different probability laws. We might expect (and it will be so) that the likelihood function proves useful in distinguishing between the different possible probability laws that could have generated the same sample values. One point to understand clearly is that we will generally be unable to "prove" a hypothesis is either true or false, in the deductive sense, based on observed values of random variables. Again, referring to the life-testing example, even if we observe sample values that yield $\bar{x}/\bar{y} = 4$, it is quite possible that $\lambda_2 = 4.1\lambda_1$, or $\lambda_2 = 3.8\lambda_1$, and so on. Observing $\bar{x}/\bar{y} = 4$ is a long way from "proving" that $\lambda_2 = 4\lambda_1$. We will use the standard terminology of *accepting* the hypothesis (if the observed sample values seem consistent with it) or *rejecting* the hypothesis (if they are not).

We will begin our discussion at the simplest possible point, one that in general is unrealistic but which allows a clear exposition of the issues involved and the statistical methodology employed to resolve them. Throughout this chapter we will assume the *form* of the probability laws generating our random samples is known. For example, our observed times to failure are a random sample of an exponential random variable, or the spectrometric readings we observe are a random sample of a normal random variable, and so on. What we do not know (which is the concern of the hypothesis to be tested) is the value(s) of the parameter(s) of the probability law. For this reason the techniques we will discuss are frequently called *parametric* tests.

8.1 Simple Hypotheses

The simplest possible situation is one in which we have observed a random sample, $X_1, X_2, \ldots, X_n$, of a random variable X and want to choose between two distinct, completely specified probability laws for X. A hypothesis that completely specifies the probability law for X is called *simple*.

Definition 8.1.1. A *simple hypothesis* H is any statement that completely specifies the probability law for a random variable X. A hypothesis that is not simple is called *composite*. A *test* of a hypothesis, H, is any rule that tells us whether to accept H or reject H, for every possible observed random sample of X. ■

For example, if we have found $n = 4$ observed spectrometer readings to be 61, 47, 53, 58, and assume these are the observed values of a random sample of a random variable X, then

H: X is normal, $\mu = 60, \sigma = 5$
H: X is normal, $\mu = 50, \sigma = 10$
H: X is normal, $\mu = 100, \sigma = 2$
H: X is exponential, $\lambda = .02$

are each simple hypotheses. Any statement, such as, "X is normal, $\mu = 60$" or "X is normal, $\sigma = 5$" or "X is exponential," which does not *completely* specify the probability law for X is a composite, not a simple, hypothesis. We will consider only simple hypotheses in this section.

In the straightforward (admittedly unrealistic) case of deciding between two simple hypotheses, it is fairly easy to find the best possible *test* (rule for deciding which of the two hypotheses should be accepted). To distinguish between the two hypotheses considered, we will call one of them the *null hypothesis*, denoted by H_0, and the other, the *alternative hypothesis*, denoted by H_1. When we test a simple H_0 versus a simple H_1, our rule must, for any possible observed sample values, tell us which of the two hypotheses to accept; thus "accept H_0" is equivalent to "reject H_1" and vice versa. To avoid confusion, we will always apply the terminology accept–reject to the *null* hypothesis.

A little thought immediately reveals that we could make two different possible errors in testing a simple H_0 versus a simple H_1. These are called "type I error" and "type II error," as defined in Table 8.1.1. Any test of H_0 will tell us *either* to accept H_0 or reject H_0, based on the observed sample values. Thus it is not possible to commit both errors simulfor any given problem. We will seek to find the rule that may be best in some sense. For example, in the life-testing case, recall that $\overline{X}$ estimates $1/\lambda_1$ (the mean time to failure, standard use) whereas $\overline{Y}$ estimates $1/\lambda_2$ (mean time to failure in the severe environment). If in actual fact it is true that $\lambda_2 = 4\lambda_1$, then we would expect to find $\overline{X}/\overline{Y}$ equal to approximately $(1/\lambda_1)/(1/4\lambda_1) = 4$. But since both $\overline{X}$ and $\overline{Y}$ are (continuous) random variables, and both would vary from one collection of possible observed values to another, we realize that we will not find $\bar{x}/\bar{y} = 4$ for all possible samples, even if $\lambda_2 = 4\lambda_1$. Thus the thought that might immediately occur, that we accept the hypothesis $\lambda_2 = 4\lambda_1$, only if we find $\bar{x}/\bar{y} = 4$, really is not totally practical. The variability in both $\overline{X}$ and $\overline{Y}$ makes a range of values for $\bar{x}/\bar{y}$ possible, when $\lambda_2 = 4\lambda_1$; thus we probably should settle on some rule (test) like "Accept the hypothesis $\lambda_2 = 4\lambda_1$ if $\bar{x}/\bar{y}$ is close to 4," where it remains to be seen what might be meant by the phrase "close to 4." As we will see, this is in fact the best type of rule from several points of view.

The observed values of random samples contain information about the probability law, but they do not contain "perfect" information, because the same set of observed values could come from many different probability laws. We might expect (and it will be so) that the likelihood function proves useful in distinguishing between the different possible probability laws that could have generated the same sample values. One point to understand clearly is that we will generally be unable to "prove" a hypothesis is either true or false, in the deductive sense, based on observed values of random variables. Again, referring to the life-testing example, even if we observe sample values that yield $\bar{x}/\bar{y} = 4$, it is quite possible that $\lambda_2 = 4.1\lambda_1$, or $\lambda_2 = 3.8\lambda_1$, and so on. Observing $\bar{x}/\bar{y} = 4$ is a long way from "proving" that $\lambda_2 = 4\lambda_1$. We will use the standard terminology of *accepting* the hypothesis (if the observed sample values seem consistent with it) or *rejecting* the hypothesis (if they are not).

We will begin our discussion at the simplest possible point, one that in general is unrealistic but which allows a clear exposition of the issues involved and the statistical methodology employed to resolve them. Throughout this chapter we will assume the *form* of the probability laws generating our random samples is known. For example, our observed times to failure are a random sample of an exponential random variable, or the spectrometric readings we observe are a random sample of a normal random variable, and so on. What we do not know (which is the concern of the hypothesis to be tested) is the value(s) of the parameter(s) of the probability law. For this reason the techniques we will discuss are frequently called *parametric* tests.

8.1 Simple Hypotheses

The simplest possible situation is one in which we have observed a random sample, $X_1, X_2, \ldots, X_n$, of a random variable X and want to choose between two distinct, completely specified probability laws for X. A hypothesis that completely specifies the probability law for X is called *simple.*

DEFINITION 8.1.1. A *simple hypothesis* H is any statement that completely specifies the probability law for a random variable X. A hypothesis that is not simple is called *composite.* A *test* of a hypothesis, H, is any rule that tells us whether to accept H or reject H, for every possible observed random sample of X. ■

For example, if we have found $n = 4$ observed spectrometer readings to be 61, 47, 53, 58, and assume these are the observed values of a random sample of a random variable X, then

$$\begin{aligned}
&H\colon\ X \text{ is normal}, \mu = 60, \sigma = 5\\
&H\colon\ X \text{ is normal}, \mu = 50, \sigma = 10\\
&H\colon\ X \text{ is normal}, \mu = 100, \sigma = 2\\
&H\colon\ X \text{ is exponential}, \lambda = .02
\end{aligned}$$

are each simple hypotheses. Any statement, such as, "X is normal, $\mu = 60$" or "X is normal, $\sigma = 5$" or "X is exponential," which does not *completely* specify the probability law for X is a composite, not a simple, hypothesis. We will consider only simple hypotheses in this section.

In the straightforward (admittedly unrealistic) case of deciding between two simple hypotheses, it is fairly easy to find the best possible *test* (rule for deciding which of the two hypotheses should be accepted). To distinguish between the two hypotheses considered, we will call one of them the *null hypothesis*, denoted by H_0, and the other, the *alternative hypothesis*, denoted by H_1. When we test a simple H_0 versus a simple H_1, our rule must, for any possible observed sample values, tell us which of the two hypotheses to accept; thus "accept H_0" is equivalent to "reject H_1" and vice versa. To avoid confusion, we will always apply the terminology accept–reject to the *null* hypothesis.

A little thought immediately reveals that we could make two different possible errors in testing a simple H_0 versus a simple H_1. These are called "type I error" and "type II error," as defined in Table 8.1.1. Any test of H_0 will tell us *either* to accept H_0 or reject H_0, based on the observed sample values. Thus it is not possible to commit both errors simulfor any given problem. We will seek to find the rule that may be best in some sense. For example, in the life-testing case, recall that $\bar{X}$ estimates $1/\lambda_1$ (the mean time to failure, standard use) whereas $\bar{Y}$ estimates $1/\lambda_2$ (mean time to failure in the severe environment). If in actual fact it is true that $\lambda_2 = 4\lambda_1$, then we would expect to find $\bar{X}/\bar{Y}$ equal to approximately $(1/\lambda_1)/(1/4\lambda_1) = 4$. But since both $\bar{X}$ and $\bar{Y}$ are (continuous) random variables, and both would vary from one collection of possible observed values to another, we realize that we will not find $\bar{x}/\bar{y} = 4$ for all possible samples, even if $\lambda_2 = 4\lambda_1$. Thus the thought that might immediately occur, that we accept the hypothesis $\lambda_2 = 4\lambda_1$, only if we find $\bar{x}/\bar{y} = 4$, really is not totally practical. The variability in both $\bar{X}$ and $\bar{Y}$ makes a range of values for $\bar{x}/\bar{y}$ possible, when $\lambda_2 = 4\lambda_1$; thus we probably should settle on some rule (test) like "Accept the hypothesis $\lambda_2 = 4\lambda_1$ if $\bar{x}/\bar{y}$ is close to 4," where it remains to be seen what might be meant by the phrase "close to 4." As we will see, this is in fact the best type of rule from several points of view.

The observed values of random samples contain information about the probability law, but they do not contain "perfect" information, because the same set of observed values could come from many different probability laws. We might expect (and it will be so) that the likelihood function proves useful in distinguishing between the different possible probability laws that could have generated the same sample values. One point to understand clearly is that we will generally be unable to "prove" a hypothesis is either true or false, in the deductive sense, based on observed values of random variables. Again, referring to the life-testing example, even if we observe sample values that yield $\bar{x}/\bar{y} = 4$, it is quite possible that $\lambda_2 = 4.1\lambda_1$, or $\lambda_2 = 3.8\lambda_1$, and so on. Observing $\bar{x}/\bar{y} = 4$ is a long way from "proving" that $\lambda_2 = 4\lambda_1$. We will use the standard terminology of *accepting* the hypothesis (if the observed sample values seem consistent with it) or *rejecting* the hypothesis (if they are not).

We will begin our discussion at the simplest possible point, one that in general is unrealistic but which allows a clear exposition of the issues involved and the statistical methodology employed to resolve them. Throughout this chapter we will assume the *form* of the probability laws generating our random samples is known. For example, our observed times to failure are a random sample of an exponential random variable, or the spectrometric readings we observe are a random sample of a normal random variable, and so on. What we do not know (which is the concern of the hypothesis to be tested) is the value(s) of the parameter(s) of the probability law. For this reason the techniques we will discuss are frequently called *parametric* tests.

Table 8.1.1

	H_0 is true	H_0 is false
Accept H_0	No error	Type II error
Reject H_0	Type I error	No error

taneously. We will define

$$\begin{aligned}\alpha &= P(\text{type I error})\\ &= P(\text{reject } H_0, \text{ given } H_0 \text{ true})\\ &= P(\text{reject } H_0 \mid H_0 \text{ true})\\ \beta &= P(\text{type II error})\\ &= P(\text{accept } H_0 \mid H_1 \text{ true}).\end{aligned}$$

Every test of H_0 has values for the pair (α, β) associated with it. It would seem ideal if we could find the test that simultaneously minimizes both α and β, but this is not possible. Since each of α and β is a probability, we know $\alpha \geq 0$ and $\beta \geq 0$: that is, 0 is the minimum value for each. No matter what H_0 and H_1 state and what observed values occur in the sample, we could use the rule (test): Accept H_0. With this test we would never commit a type I error, since we would not reject H_0, no matter what the sample values were. Thus for this test we would have $\alpha = 0$, its smallest possible value (and this test has $\beta = 1$, its largest possible value). The converse of this test, which would always reject H_0, gives $\beta = 0$, $\alpha = 1$. Neither of these tests is desirable, because they maximize one of the two probabilities of error while minimizing the other. The following example illustrates the type of trade-off that typically exists between α and β.

EXAMPLE 8.1.1. Assume $X_1, X_2, \ldots, X_9$ is a random sample of a normal random variable whose variance σ^2 is known to equal 1, and we want to test $H_0: \mu = 2$ versus $H_1: \mu = 3$. Since the two simple hypotheses specify values for μ, and $\bar{X} = \frac{1}{9}\sum X_i$ is the minimum variance unbiased estimator for μ, it would seem reasonable to use a test that recommends acceptance or rejection of H_0, based on the observed value for $\bar{X}$. More specifically, if H_0 is true, we would expect the observed value for $\bar{X}$ to be close to 2 rather than to 3; if H_0 is not true, we would expect to find the observed value for $\bar{X}$ close to 3 rather than to 2. Thus we might consider a test that accepts H_0 if $\bar{x} \leq c$ and which rejects H_0 if $\bar{x} > c$, where c is a constant to be chosen. Figure 8.1.1 graphs the density for $\bar{X}$ for the two

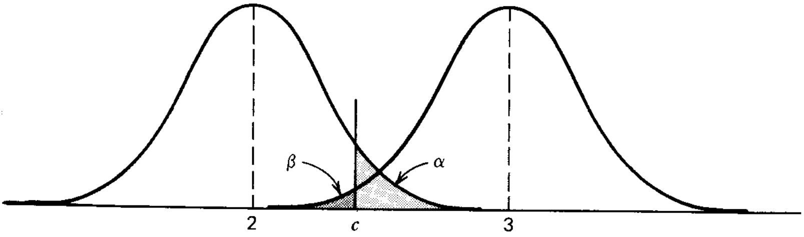

Figure 8.1.1

the cases and indicates the values of α and β for a particular c. We have

$$\begin{aligned}\alpha &= P(\text{reject } H_0 | H_0 \text{ true})\\ &= P(\bar{X} > c | H_0 \text{ true})\\ &= 1 - N\left(\frac{c-2}{1/3}\right)\\ &= N(3(2-c))\end{aligned}$$

X is normal with mean 2, variance $\frac{1}{9}$ if H_0 is true. Similarly,

$$\begin{aligned}\beta &= P(\text{accept } H_0 | H_0 \text{ false})\\ &= P(\bar{X} \le c | H_0 \text{ false})\\ &= N\left(\frac{c-3}{1/3}\right) = N(3(c-3))\end{aligned}$$

X is normal with mean 3, variance $\frac{1}{9}$ if H_0 is false (meaning H_1 is true). 8.1.2 gives four choices for c, with the resulting values of α and β. make c larger, the value for α decreases but the value for β increases.

Table 8.1.2

c	α	β
2.2	.2743	.0082
2.4	.1151	.0359
2.6	.0359	.1151
2.8	.0082	.2743

■

The trade-off in magnitude of α and β, as illustrated in Table 8.1.2, is typical. What then is a reasonable procedure to use in choosing a test? In order to discuss this choice a little more exactly, let us assume we want to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where θ_0 and θ_1 are specified values for the parameter of the probability law for a discrete random variable X. We will have a random sample, $X_1, X_2, \ldots, X_n$, of X to use in making our decision to accept or reject H_0. Let us use

$$\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$$

to represent the observed sample values and let S be the sample space for $\underset{\sim}{x}$ (collection of all possible $\underset{\sim}{x}$ values that might occur). The likelihood function for the sample then has only two possible values, $L_X(\theta_0)$ or $L_X(\theta_1)$, because we are assuming that these are the only possible values for θ. Now recall that $L_X(\theta_0)$ is the product of the probability functions evaluated at the observed sample values, with θ_0 the value for the parameter. Thus $L_X(\theta_0)$ actually gives the probability of observing $\underset{\sim}{x}$, assuming θ_0 is the correct parameter value and $L_X(\theta_1)$ gives the probability of observing $\underset{\sim}{x}$, assuming θ_1 is the correct parameter value.

Any test of H_0 versus H_1 must give the decision "accept H_0" or the decision "reject H_0" for any possible observed $\underset{\sim}{x}$. Thus we can think of any test as providing a partition of S, the sample space for $\underset{\sim}{x}$, into two parts.

$$\begin{aligned} A &= \{\underset{\sim}{x} : \text{we accept } H_0\} \\ R &= \{\underset{\sim}{x} : \text{we reject } H_0\}. \end{aligned}$$

We will call A the *acceptance region* for the test and R the *rejection* (or *critical*) *region* for the test, and since they are a partition of S, $A \cup R = S$ and $A \cap R = \varnothing$ (these simply say we must make exactly one of the two decisions for every possible $\underset{\sim}{x} \in S$). Thus every test can be thought of as a partition of S or, equivalently, the class of all possible tests of H_0 versus H_1 is defined by all possible partitions of S. For the two trivial tests mentioned earlier we have

$$\begin{array}{lll} \text{Always accept } H_0: & A = S & R = \varnothing \\ \text{Always reject } H_0: & A = \varnothing & R = S, \end{array}$$

so each of these provides a (trivial) partition of S.

Since $L_X(\theta_0)$ and $L_X(\theta_1)$ each provide the probability of observing $\underset{\sim}{x}$, note that

$$\sum_{\underset{\sim}{x} \in S} L_X(\theta_0) = 1$$

$$\sum_{\underset{\sim}{x} \in S} L_X(\theta_1) = 1,$$

the sum of the probabilities of all possible observed $\underset{\sim}{x}$ must, of course, equal 1 in either case. But then for any specified test (partition of S into A, R)

$$\begin{aligned}\sum_{\underset{\sim}{x} \in S} L_X(\theta_0) &= \sum_{\underset{\sim}{x} \in A} L_X(\theta_0) + \sum_{\underset{\sim}{x} \in R} L_X(\theta_0)\\ &= P(\text{accept } H_0 | \theta_0) + P(\text{reject } H_0 | \theta_0)\\ &= (1 - \alpha) + \alpha\\ \sum_{\underset{\sim}{x} \in S} L_X(\theta_1) &= \sum_{\underset{\sim}{x} \in A} L_X(\theta_1) + \sum_{\underset{\sim}{x} \in R} L_X(\theta_1)\\ &= P(\text{accept } H_0 | \theta_1) + P(\text{reject } H_0 | \theta_1)\\ &= \beta + (1 - \beta),\end{aligned}$$

so the probabilities of the two types of error are given by summing the appropriate likelihood function over the correct region:

$$\alpha = \sum_{\underset{\sim}{x} \in R} L_X(\theta_0), \qquad \beta = \sum_{\underset{\sim}{x} \in A} L_X(\theta_1).$$

Any reasonable test should accept H_0 for any outcome $\mathbf{x}$ such that $L_X(\theta_0) > L_X(\theta_1)$, because if this inequality is true, it says this $\mathbf{x}$ is more likely to occur if θ_0 is the true parameter value than if θ_1 is the true parameter value. Let $k > 0$ be an arbitrary constant and consider the particular test that has as its acceptance region

$$A = \{\underset{\sim}{x}: L_X(\theta_0) > kL_X(\theta_1)\};$$

the rejection region for this test then is

$$R = \{\underset{\sim}{x}: L_X(\theta_0) \le kL_X(\theta_1)\}.$$

Now suppose we choose the value for k so that

$$\alpha = \sum_{\underset{\sim}{x} \in R} L_X(\theta_0)$$

is fixed (say, at .01 or .05 or whatever value we choose), and consider any other test (with partition A^*, R^*), which has the same probability of type I error, that is,

$$\alpha = \sum_{\underset{\sim}{x} \in R^*} L_X(\theta_0)$$

as well. Then the two tests together partition S into four parts: $A \cap A^*$,

$A \cap R^*$, $R \cap A^*$, $R \cap R^*$, and we know

$$\begin{aligned}0 = \alpha - \alpha &= \sum_{\underset{\sim}{x} \in R} L_X(\theta_0) - \sum_{\underset{\sim}{x} \in R^*} L_X(\theta_0)\\ &= \left(\sum_{\underset{\sim}{x} \in R \cap R^*} L_X(\theta_0) + \sum_{\underset{\sim}{x} \in R \cap A^*} L_X(\theta_0) \right)\\ &\quad - \left(\sum_{\underset{\sim}{x} \in A \cap R^*} L_X(\theta_0) + \sum_{\underset{\sim}{x} \in R \cap R^*} L_X(\theta_0) \right)\\ &= \sum_{\underset{\sim}{x} \in R \cap A^*} L_X(\theta_0) - \sum_{\underset{\sim}{x} \in A \cap R^*} L_X(\theta_0)\end{aligned}$$

so it follows that

$$\sum_{\underset{\sim}{x} \in R \cap A^*} L_X(\theta_0) = \sum_{\underset{\sim}{x} \in A \cap R^*} L_X(\theta_0).$$

But for all $\underset{\sim}{x} \in R$ (and thus those in $R \cap A^*$ in particular) we have

$$L_X(\theta_0) \leq kL_X(\theta_1)$$

and for all $\underset{\sim}{x} \in A$ (and in $A \cap R^*$)

$$L_X(\theta_0) > kL_X(\theta_1).$$

Thus

$$\sum_{\underset{\sim}{x} \in A \cap R^*} L_X(\theta_0) > k \sum_{\underset{\sim}{x} \in A \cap R^*} L_X(\theta_1)$$

and

$$- \sum_{\underset{\sim}{x} \in R \cap A^*} L_X(\theta_0) \geq -k \sum_{\underset{\sim}{x} \in R \cap A^*} L_X(\theta_1);$$

adding these two inequalities gives

$$0 \geq k \sum_{\underset{\sim}{x} \in A \cap R^*} L_X(\theta_1) - k \sum_{\underset{\sim}{x} \in R \cap A^*} L_X(\theta_1)$$

or

$$\sum_{\underset{\sim}{x} \in R \cap A^*} L_X(\theta_1) \geq \sum_{\underset{\sim}{x} \in A \cap R^*} L_X(\theta_1).$$

Adding

$$\sum_{\underset{\sim}{x} \in A \cap A^*} L_X(\theta_1)$$

to both sides gives

$$\sum_{\underset{\sim}{x} \in R \cap A^*} L_X(\theta_1) + \sum_{\underset{\sim}{x} \in A \cap A^*} L_X(\theta_1) \geq \sum_{\underset{\sim}{x} \in A \cap R^*} L_X(\theta_1) + \sum_{\underset{\sim}{x} \in A \cap A^*} L_X(\theta_1),$$

that is,

$$\sum_{\underset{\sim}{x} \in A^*} L_X(\theta_1) \geq \sum_{\underset{\sim}{x} \in A} L_X(\theta_1).$$

This equation says the probability of accepting H_0 when $\theta = \theta_1$, using the test with partition (A^*, R^*), must be at least as large as the probability of accepting H_0 (again, when $\theta = \theta_1$), using the test with A, R as previously defined. That is, if we consider any other test with the same probability of type I error, α, its probability of type II error must be at least as large as β for the test with rejection region

$$R = \{\mathbf{x}\colon L_X(\theta_0) \leq kL_X(\theta_1)\}$$

so the test with this rejection region minimizes β for α fixed. This establishes the following theorem, named after J. Neyman and E. Pearson who first published it in the 1930s. It forms the basis for choosing the "best" test of a simple H_0 versus a simple H_1.

Theorem 8.1.1. To test the simple $H_0\colon \theta = \theta_0$ versus the simple $H_1\colon \theta = \theta_1$, based on a random sample of size n of a random variable whose probability law depends on θ, the test with critical region

$$R = \{\underset{\sim}{x}\colon L_X(\theta_0) \leq kL_X(\theta_1)\}$$

has the smallest possible value for

$$\beta = \sum_{\underset{\sim}{x} \in A} L_X(\theta_1)$$

among all tests with the same value for

$$\alpha = \sum_{\underset{\sim}{x} \in R} L_X(\theta_0). \qquad ■$$

This theorem gives a very simple way of finding the partition of S that will give the smallest possible probability of type II error, for any chosen value for α. One simply uses the likelihood function to assign sample values to the rejection (or critical) region R and adjusts the value for k to give the desired value for α. Exactly the same reasoning can be employed with samples of continuous random variables, using integration rather than summation. Thus we again define the critical region by comparing the values of the likelihood function and adjust k to give the desired value for α.

For all the standard probability laws, this best partitioning of S, the sample space for $\underset{\sim}{x}$, reduces to an equivalent partition of the possible range for some statistic $g(\underset{\sim}{x})$, a function of the observed values. Thus the test can equally well, and more simply, be expressed in terms of $g(\underset{\sim}{x})$, the observed value for $g(\underset{\sim}{X})$. We will refer to $g(\underset{\sim}{X})$ as being the *test statistic* for the given hypotheses. The following examples apply this theorem to some standard cases.

EXAMPLE 8.1.2. Let us reconsider the case discussed in Example 8.1.1 and contrast what we did there to construct a test of $H_0: \mu = 2$ versus $H_1: \mu = 3$, with this best procedure. Given a random sample of size n of a normal random variable with mean μ and known $\sigma^2 = 1$ (so the hypotheses are simple), the two values for the likelihood function are

$$L_X(2) = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left[\frac{-\sum(x_i - 2)^2}{2}\right]$$

$$L_X(3) = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left[\frac{-\sum(x_i - 3)^2}{2}\right].$$

The best test then has critical region

$$R = \left\{\underset{\sim}{x}: \left(\frac{1}{2\pi}\right)^{n/2} \exp\left[\frac{-\sum(x_i - 2)^2}{2}\right] \le k\left(\frac{1}{2\pi}\right)^{n/2} \exp\left[\frac{-\sum(x_i - 3)^2}{2}\right]\right\}$$

which is equivalent to

$$R = \{\underset{\sim}{x}: \sum(x_i - 2)^2 \ge -2\ln k + \sum(x_i - 3)^2\}$$

and to

$$R = \left\{\underset{\sim}{x}: \sum x_i \ge -\ln k + \frac{5n}{2}\right\}$$

and to

$$R = \left\{\underset{\sim}{x}: \bar{x} \ge -\frac{1}{n}\ln k + \frac{5}{2} = c\right\},$$

the same critical region we employed. Thus $\bar{X}$ is called the test statistic (since only its value is needed to decide whether to accept or reject H_0) and we can choose c to set α at any value we like (you can verify that $c = 2 + \frac{1}{3}z_{1-\alpha}$). There is no real interest in finding the value for k (although we could if we want), because it is much simpler to compare the observed value for $\bar{X}$ with c rather than evaluating the likelihood functions, $L_X(2)$ and $L_X(3)$ and checking to see whether $L_X(2) \leq kL_X(3)$. ■

EXAMPLE 8.1.3. Suppose we assume the time to failure X for a piece of equipment is an exponential random variable with parameter λ. Given a random sample $X_1, X_2, \ldots, X_n$ of lifetimes we want to test $H_0: \lambda = .01$ versus $H_1: \lambda = .04$. The likelihood functions for the two cases then are

$$\begin{aligned} L_X(.01) &= (.01)^n e^{-.01\,\Sigma x_i} \\ L_X(.04) &= (.04)^n e^{-.04\,\Sigma x_i} \end{aligned}$$

and the best critical region is defined by

$$R = \{\underset{\sim}{x}: (.01)^n e^{-.01\,\Sigma x_i} \leq k(.04)^n e^{-.04\,\Sigma x_i}\},$$

which you can verify is equivalent to

$$R = \{\underset{\sim}{x}: \bar{x} \leq c\}$$

so again, $\bar{X}$ is the test statistic. Since if $H_0: \lambda = .01$ is true, $2(.01)\,n\bar{X}$ is a χ^2 random variable with $2n$ degrees of freedom, to have any desired value for α we need

$$\begin{aligned} \alpha &= P(\bar{X} \leq c \mid H_0 \text{ true}) \\ &= P(2(.01)\,n\bar{X} \leq (.02)\,nc \mid H_0 \text{ true}) \\ &= P(\chi^2 \leq (.02)\,nc) \end{aligned}$$

so

$$\chi^2_\alpha = .02nc$$

and we use

$$c = \frac{50\chi^2_\alpha}{n}.$$

Thus if we take a sample of $n = 8$ lifetimes and want $\alpha = .1$, we find (for $2n = 16$ degrees of freedom),

$$\chi^2_{.1} = 9.31, \qquad c = \frac{(50)(9.31)}{8} = 58.19$$

and we should reject H_0 if we find $\bar{x} \leq 58.19$. This test has the smallest possible β among all those with $\alpha = .1$. ■

We can, of course, also employ this procedure to find the best test for simple hypotheses about the parameter of a discrete probability law. Since the test statistic then is in general a function of the discrete sample values, it is itself a discrete random variable. Because of this discreteness, only a discrete collection of values for α are attainable. Nonetheless, if we select k to have any attainable α value, the Neyman–Pearson critical region still gives the test with the smallest possible β among all those with the same α (or any smaller α, actually).

EXAMPLE 8.1.4. Assume a large lot of items is received, each of which is either defective or nondefective; the lot is large enough so that it is reasonable to assume individual items selected at random without replacement are independent Bernoulli trials with parameter p, where p is the proportion of defectives in the lot. Suppose $n = 50$ items are selected at random, tested, and $\sum_{i=1}^{50} X_i$ is the number of defectives found (in the n tested); we want to test H_0: $p = .1$ versus H_1: $p = .2$. The two likelihood functions then are

$$L_X(.1) = (.1)^{\Sigma x_i}(.9)^{50-\Sigma x_i}$$
$$L_X(.2) = (.2)^{\Sigma x_i}(.8)^{50-\Sigma x_i}$$

and the best critical region is

$$R = \{\mathbf{x}: (.1)^{\Sigma x_i}(.9)^{50-\Sigma x_i} \le k(.2)^{\Sigma x_i}(.8)^{50-\Sigma x_i}\},$$

which is equivalent to

$$R = \{\underset{\sim}{x}: \sum x_i \ge c\},$$

so $Y = \sum_{i=1}^{50} X_i$ is the test statistic. Y is a binomial random variable with parameters 50 and p and, if H_0 is true,

$$P(Y \ge c \mid p = .1) = \sum_{j=c}^{50} \binom{50}{j} (.1)^j (.9)^{50-j}$$

so the possible values for α are limited by the values that can be achieved by this discrete sum. The values given in Table 8.1.3 can be verified without difficulty on a hand-held calculator.

Table 8.1.3

c	$P(Y \ge c \mid p = .10)$
7	.230
8	.122
9	.058
10	.025
11	.009

Thus with a sample of $n = 50$, the best test of H_0: $p = .1$ versus H_1: $p = .2$ says we should reject H_0 if $y \geq 9$, if we want $\alpha = .058$, and reject H_0 if $y \geq 10$, if we want $\alpha = .025$. In either case we are using the test with the smallest β among all those whose probability of type I error does not exceed the α value chosen. The choice of which α value to use, of course, depends on the person applying the test and her desired probability of committing a type I error. ■

There is in actual fact a rather close connection between tests of hypotheses and estimation of parameters. In general, any statistic that provides a good estimator for an unknown parameter will also be the test statistic for testing hypotheses about the same parameter. The form of the best critical region (as defined by Theorem 8.1.1) can also generally be surmised by considering the values of the estimator (test statistic) that are more consistent with H_1 than with H_0.

In testing a simple H_0 versus a simple H_1, we see from Theorem 8.1.1 how, for a fixed sample size n, to find the test (critical region) that gives the smallest value for β for any given value of α. There are three basic quantities involved in such a test: the sample size n, α, and β. In testing simple hypotheses about the mean of a normal distribution, we can select, quite straightforwardly, desired values for α and β, and then find the sample size n that, together with the best test (partition of all possible observed samples), gives the desired α and β. This is illustrated in the following example.

EXAMPLE 8.1.5. Assume a normal random variable X has unknown mean μ and known variance σ^2 and we want to test H_0: $\mu = \mu_0$ versus H_1: $\mu = \mu_1$, where μ_0 and $\mu_1 > \mu_0$ are any desired constants. Further suppose we would like to fix both α and β, the probabilities of the two types of error. How large should n, the sample size, be? As we know from Example 8.1.2, the best critical region is one in which we reject H_0 if $\bar{x} \geq c$, as long as $\mu_1 > \mu_0$ (if $\mu_1 < \mu_0$, the best critical region is specified by $\bar{x} \leq c$). Given a random sample of size n of X, we know that $\bar{X}$ is normal with mean μ_0, variance σ^2/n if H_0 is true, and $\bar{X}$ is normal with mean μ_1, variance σ^2/n if H_1 is true. Thus with $\mu_1 > \mu_0$,

$$\begin{aligned}\alpha &= P(\text{type I error})\\ &= P(\bar{X} \geq c \mid \mu = \mu_0)\\ &= P\left(\frac{(\bar{X} - \mu_0)\sqrt{n}}{\sigma} \geq \frac{(c - \mu_0)\sqrt{n}}{\sigma} \,\middle|\, \mu = \mu_0\right)\end{aligned}$$

$$= 1 - N\left(\frac{(c - \mu_0)\sqrt{n}}{\sigma}\right)$$
$$= N\left(\frac{\sqrt{n}(\mu_0 - c)}{\sigma}\right)$$

so we require

$$\frac{\sqrt{n}(\mu_0 - c)}{\sigma} = z_\alpha.$$

Similarly, the probability of a type II error is

$$\beta = P(\overline{X} \le c \,|\, \mu = \mu_1)$$
$$= P\left(\frac{(\overline{X} - \mu_1)\sqrt{n}}{\sigma} \le \frac{(c - \mu_1)\sqrt{n}}{\sigma} \,\middle|\, \mu = \mu_1\right)$$
$$= N\left(\frac{(c - \mu_1)\sqrt{n}}{\sigma}\right)$$

so we also need

$$\frac{\sqrt{n}(c - \mu_1)}{\sigma} = z_\beta.$$

Thus we have two equations

$$\mu_0 - c = z_\alpha \frac{\sigma}{\sqrt{n}}$$
$$c - \mu_1 = z_\beta \frac{\sigma}{\sqrt{n}}$$

in two unknowns, c and $\sqrt{n}$. It is easy to see that the solutions are given by

$$\sqrt{n} = \frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu_1}, \qquad c = \frac{z_\alpha \mu_1 + z_\beta \mu_0}{z_\alpha + z_\beta},$$

so a sample of size

$$n = \frac{\sigma^2 (z_\alpha + z_\beta)^2}{(\mu_0 - \mu_1)^2},$$

employed with the best test, gives the desired values for α and β (this value for n is not likely to be an integer; the conservative approach then is to round up to the next higher integer).

For example, suppose we assume X is normal with $\sigma = 2$ and we would like to find the sample size n that, using the best test, will give $\alpha = .01$, $\beta = .05$ in testing H_0: $\mu = 2$ versus H_1: $\mu = 5$. We have $z_\alpha = z_{.01} = -2.33$, $z_\beta = z_{.05} = -1.64$, so the preceding solution gives

$$n = \frac{2^2(-2.33 - 1.64)^2}{(2 - 5)^2} = 7.005$$

so a sample of $n = 8$ will be sufficient. With $n = 8$, if we want to hold $\alpha = .01$, we find c from

$$\frac{\sqrt{8}(2 - c)}{2} = z_{.01} = -2.33,$$

which gives

$$c = 2 + \frac{4.66}{\sqrt{8}} = 3.65.$$

The value for β then would be

$$P(\bar{X} \leq 3.65 | \mu = 5) = N\left(\frac{(3.65 - 5)\sqrt{8}}{2}\right)$$
$$= N(-1.91) = .0281,$$

smaller than the desired $\beta = .05$ because we rounded n up to 8. On the other hand, we could keep $\beta = .05$ by using the value for c determined from

$$\frac{\sqrt{8}(c - 5)}{2} = z_\beta = -1.64,$$

which gives

$$c = 5 - \frac{3.28}{\sqrt{8}} = 3.84;$$

the value for α then is

$$P(\bar{X} \geq 3.84 | \mu = 2) = 1 - N\left(\frac{(3.84 - 2)\sqrt{8}}{2}\right)$$
$$= 1 - N(2.60) = .0047.$$

If we round n up, as previously recommended, either α or β, or both, will be smaller than the values initially specified. ■

It is, of course, extremely rare that one would have a random sample from a normal population whose variance σ^2 is known but whose mean μ is unknown. Thus the procedure discussed in the preceding example is not frequently employed per se. It is, however, a useful prototype for cases in which the test statistic, used in the best test of a simple H_0 versus a simple H_1, has a distribution that is well approximated by the normal for large n. This is illustrated in the following example.

EXAMPLE 8.1.6. Assume, as in Example 8.1.4, a large lot of items contains the proportion p of defectives; we want to test H_0: $p = .1$ versus H_1: $p = .2$ with $\alpha = \beta = .01$, say (at least approximately). How large should the sample size n be? Again, as long as the number of items in the lot is large enough that inspection of individual items can be well approximated by independent Bernoulli trials, each with parameter p, the number of defectives Y in a random sample of size n is binomial with parameters n and p. For large n, Y is approximately normal, $\mu = np$, $\sigma^2 = npq$, so to have $\alpha = .01$, we want

$$\begin{aligned} .01 &= P(Y \geq c \mid p = .1) \\ &= 1 - P(Y \leq c - 1 \mid p = .1) \\ &\doteq 1 - N\left(\frac{(c - 1 - .1n)}{\sqrt{n(.1)(.9)}}\right) \end{aligned}$$

which says then we want

$$\frac{c - 1 - .1n}{\sqrt{.09n}} \doteq 2.33.$$

To have $\beta = .01$, we want

$$\begin{aligned} .01 &= P(Y \leq c - 1 \mid p = .2) \\ &\doteq N\left(\frac{(c - 1 - .2n)}{\sqrt{n(.2)(.8)}}\right) \end{aligned}$$

so we also require

$$\frac{c - 1 - .2n}{\sqrt{.16n}} \doteq -2.33.$$

Solving these two equations simultaneously gives $n = (7 \times 2.33)^2 = 266.02$, so a sample of $n = 267$ would be sufficient; again, because we have rounded n up to the next larger integer, we can take α or β (or both) smaller than

.01. Since n has turned out to be "large," the normal approximation used should be very accurate and, with $n = 267$, we can hold both α and β to being roughly .01 in testing H_0: $p = .1$ versus H_1: $p = .2$. ■

EXERCISE 8.1

1. Suppose that X is a Bernoulli random variable with parameter p. We take a random sample of four observations of X and want to test H_0: $p = \frac{1}{4}$ versus H_1: $p = \frac{3}{4}$. If we reject H_0 only if we get four successes in the sample, compute the values of α and β.
2. Given that X is a uniform random variable on the interval $(0, \theta)$, we might test H_0: $\theta = 1$ versus the alternative H_1: $\theta = 2$ by taking a sample of 2 observations of X and rejecting H_0 if $\bar{X} > .99$. Compute α and β for this test.
3. Assume we have a random sample of size n of a continuous random variable X and want to test the simple hypothesis that the density for X is

$$f_X(x) = 2x, \qquad 0 < x < 1$$

versus the simple alternative that the density for X is

$$f_X(x) = 1, \qquad 0 < x < 1.$$

Find the best test for this hypothesis.

4. Five oil samples are removed from the same oil reservoir and analyzed on a spectrometer for their iron content. The iron readings produced are assumed to be normal with mean μ (unknown true iron contamination in the reservoir) and variance $\sigma^2 = 3$. Granted

$$\sum_{i=1}^{5} x_i = 265,$$

would you accept H_0: $\mu = 50$, with $\alpha = .05$, in testing versus H_1: $\mu = 55$? if the alternative were H_1: $\mu = 45$?

5. A random variable X is known to be normal with $\mu = 5$, σ^2 unknown. What is the test statistic and best critical region in testing H_0: $\sigma^2 = 10$ versus H_1: $\sigma^2 = 20$ for a sample of size n? How does your answer change if the alternative is H_1: $\sigma^2 = 5$?
6. The times for an auto repair shop to diagnose and repair a certain problem are assumed to be exponential with parameter λ and mean $\mu = 1/\lambda$, with units of hours. Six cars with the same problem required 1.8, 5.2, 0.4, 5.1, 0.6, 3.5 hours, respectively, to be successfully re-

paired. Based on these observed values would you accept H_0: $\mu = 4$ with $\alpha = .1$, with the alternative H_1: $\mu = 3$?

7. Assume X is a Poisson random variable with parameter μ. What is the test statistic and best critical region for testing H_0: $\mu = 2$ versus H_1: $\mu = 1$, based on a random sample of n observations of X?

8. If meteorites strike the surface of the moon "at random," the number of meteorite craters per unit of area should be a Poisson random variable with parameter μ (expected number per unit area). If $n = 10$ units of area are examined and the total number of meteorite craters found is 8, would you accept H_0: $\mu = \frac{1}{2}$, with $\alpha = .068$, with the alternative H_1: $\mu = 1$? Evaluate β for this test.

*9. Theorem 8.1.1 can also be used to find the best test that a random variable X has one completely specified probability law versus another. Based on a random sample of n observations of a continuous random variable, describe the best critical region for testing H_0: X is normal, $\mu = 100$, $\sigma = 20$ versus H_1: X is exponential, $\lambda = .01$.

10. Use the large sample (normal approximation) methodology to evaluate how large a sample, n, is required to test H_0: $\mu = 150$ versus H_1: $\mu = 200$, with $\alpha = .01$, $\beta = .02$, assuming the random variable is exponential with mean μ. Would you accept or reject H_0 if for the sample size determined, you found $\bar{x} = 175$?

11. Repeat Exercise 8.1.10 if the random variable is Poisson with parameter μ, testing H_0: $\mu = 2$ versus H_1: $\mu = 1.5$, with $\alpha = .01$, $\beta = .02$. Would you accept or reject H_0 if you found $\bar{x} = 1.75$?

12. Each of $n = 10$ persons used the same instrument to measure the same object; the true value was then subtracted from each of these. These 10 differences are assumed to be a random sample of a normal random variable X with mean 0 and variance σ^2 and are used to test H_0: $\sigma^2 = 2$ versus H_1: $\sigma^2 = 4$. The critical region used for the test is

$$R = \{\underset{\sim}{x} : \sum x_i^2 \geq 32\}.$$

Evaluate α and β for this test.

13. To test H_0: $p = .01$ versus H_1: $p = .005$, $n = 1000$ flashbulbs will be ignited, where p is the probability a flashbulb is defective (and will not flash). The rejection region is

$$R = \{\underset{\sim}{x} : \sum x_i \leq 6\},$$

where $x_i = 0$ if the ith bulb flashes, $x_i = 1$ if it does not. Evaluate α and β for this test.

8.2 Composite Hypotheses

In most applications unknown parameters of probability laws are not restricted to either of just two possible values. The probability of selecting a defective item from a large lot is not restricted to $p = .1$ or $p = .2$; indeed, if the lot contains N items, the proportion defective must be one of the values $0, 1/N, 2/N, \ldots, (N - 1)/N, 1$, which if we idealize to Bernoulli trials, actually is represented by the interval $[0, 1]$. The possible parameter values for a normal probability law are $\{(\mu, \sigma)\colon -\infty < \mu < \infty, \sigma > 0\}$, a set that is continuous in two dimensions. Thus in most applications one is interested in cases in which H_0 or H_1 (or both) is a composite, not simple, hypothesis. We will discuss tests of composite hypotheses in this section.

To keep our discussion concrete, let us first describe the case in which we have a random sample of a random variable X, whose probability law depends on a single unknown parameter θ, which can equal any value in a continuous interval on the real line (for example, X is Bernoulli and its parameter p lies in the interval $[0, 1]$, or X is exponential and its parameter λ is positive). The collection of possible values for the unknown parameter will be called the *parameter space* and will be denoted by Ω; for a Bernoulli X, the parameter space is $\Omega = \{p\colon 0 \le p \le 1\}$ and for X exponential the parameter space is $\Omega = \{\lambda\colon \lambda > 0\}$. A probability law with one unknown parameter then is said to have a one-dimensional parameter space Ω.

One of the most frequently occurring composite-hypothesis situations is referred to as testing a one-sided alternative. If θ_0 is any known, fixed value in Ω, and we want to test $H_0\colon \theta \le \theta_0$ versus the alternative $H_1\colon \theta > \theta_0$, we have a one-sided alternative; similarly, in testing $H_0\colon \theta \ge \theta_0$ versus $H_1\colon \theta < \theta_0$, we again have a one-sided alternative. In testing one-sided alternatives, we generally find a range of values for θ for which H_0 is true, as well as a range of values for θ for which H_1 is true. Thus if we assume H_0 true, we do not have a single unique probability law for the random variable X; the same is true for H_1. This contrasts with the simple versus simple case, in which assuming either H_0 or H_1 true does uniquely specify the probability law for X. The real impact of H_0 and H_1 being composite is that we no longer have single specific numbers that we can call the probabilities of type I and type II errors. When H_0 is composite, $P(\text{reject } H_0 | H_0$ true) depends on which particular value, of all those specified by H_0, we assume to be the true value for θ; actually, $P(\text{reject } H_0 | H_0 \text{ true})$ has now become a function of θ, defined for all θ specified by H_0. Similarly, with H_1 composite, $P(\text{accept } H_0 | H_1 \text{ true})$ is actually a function defined for all θ specified by H_1. Both these functions can be evaluated from the *operating characteristic* (OC) function or the *power function* of the test, defined as follows.

Definition 8.2.1. The *operating characteristic* function (frequently called the OC curve) of a test of a hypothesis H_0 about a parameter θ is

$$C(\theta) = P(\text{accept } H_0 | \theta).$$

The complementary function

$$Q(\theta) = 1 - C(\theta)$$

is called the *power function* of the test. ■

Different tests (rules for accepting or rejecting H_0) have different OC curves or, equivalently, different power functions. Because the value of an OC curve is a probability for any θ, we know immediately that $0 \le C(\theta) \le 1$ for all θ; similarly, $0 \le Q(\theta) \le 1$, the power function is also bounded by 0 and 1.

Suppose $\Omega = \{\theta : \theta > 0\}$ and we want to test $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$, where θ_0 is some specific value in Ω. The ideal OC curve (pictured in Figure 8.2.1) then would be

$$\left.\begin{aligned} C(\theta) &= 1, \quad & \theta \le \theta_0 \\ &= 0, \quad & \theta > \theta_0. \end{aligned}\right)$$

If θ is really unknown, and all we can do is select a random sample of values for X on which to base our test, this ideal OC curve can not be realized. The following examples illustrate the computation of OC curves.

Example 8.2.1. Suppose we have a random sample of size 10 of an exponential random variable X and want to test $H_0 : \lambda \le 1$ versus $H_1 : \lambda > 1$.

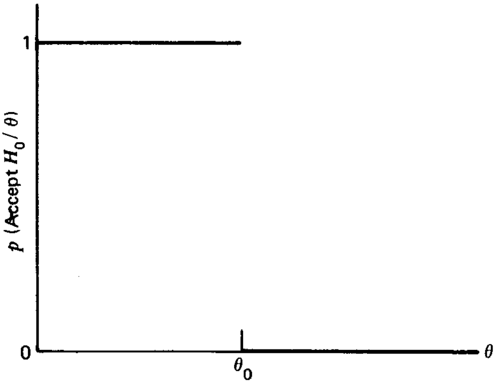

Figure 8.2.1

The critical region for our test is (rather arbitrarily)

$$R = \{\underset{\sim}{x}: \bar{x} \le .545\}.$$

The OC curve for this test then is

$$\begin{aligned} C(\lambda) &= P(\bar{X} > .545 \mid \lambda) \\ &= P(20\lambda\bar{X} > 10.9\lambda) \\ &= P(\chi^2(20) > 10.9\lambda) \\ &= 1 - F_{\chi^2}(10.9\lambda), \end{aligned}$$

one minus the χ^2 distribution function, 20 degrees of freedom, evaluated at 10.9λ, since $2\lambda n\bar{X}$ is a χ^2 random variable with $2n$ degrees of freedom. From Table 2 in the Appendix, we can evaluate certain quantiles of the χ^2 distribution function with 20 degrees of freedom and, from these, the OC curve for this test. For example, we find $F_{\chi^2}(8.26) = .01$, $10.9\lambda = 8.26$ gives $\lambda = .758$ so

$$C(.758) = 1 - F_{\chi^2}(8.26) = .99.$$

Similarly,

$$\begin{aligned} C(1) &= 1 - F_{\chi^2}(10.9) = .95 \\ C(1.771) &= 1 - F_{\chi^2}(19.3) = .5 \\ C(2.881) &= 1 - F_{\chi^2}(31.4) = .05, \text{ and so on.} \end{aligned}$$

We can directly use the χ^2 quantiles to evaluate points on the OC curve for this test. Since the $\chi^2(20)$ distribution function is monotonic increasing, the OC curve is monotonic decreasing. The OC curve for this test is pictured in Figure 8.2.2; this shape is typical for the commonly used tests. Notice that for all $\lambda \le 1$ (those values specified by H_0) we have $C(\lambda) \ge .95$ and thus

$$\begin{aligned} Q(\lambda) &= 1 - C(\lambda) \\ &= P(\text{reject } H_0 \mid \lambda) \\ &\le .05 \qquad \text{for} \qquad \lambda \le 1. \end{aligned}$$

Thus $Q(\lambda) = 1 - C(\lambda) = P(\text{reject } H_0 \mid \lambda)$ is, for λ specified by H_0, actually, the function whose values give the probability of type I error. Similarly,

$$C(\lambda) = P(\text{accept } H_0 \mid \lambda) = 1 - Q(\lambda)$$

is, for λ specified by H_1, the function whose values give the probability of type II error. The largest value of $Q(\lambda)$, for all λ specified by H_0, is called the *size* of the test and is denoted by α. Thus the size of this test is $\alpha = .05$. ■

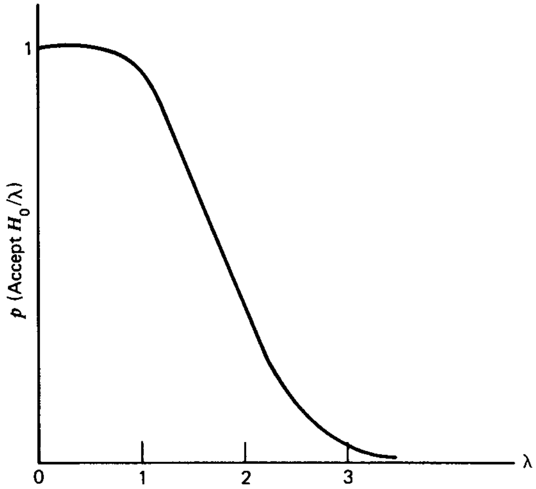

Figure 8.2.2

EXAMPLE 8.2.2. A flashbulb manufacturer claims that p, the probability any one of its bulbs is defective and will not work, is no larger than .01. Suppose we formally wish to test this claim. We want to test H_0: $p \leq .01$ versus H_1: $p > .01$; to do so we will fire 10 of the bulbs. If X, the number that do not work, is 1 or more we will reject H_0; otherwise (if $X = 0$), we will accept H_0. Granted there is a constant probability p that any one of the bulbs will not fire, and the bulbs used are independent trials, X is binomial, $n = 10$, p. The OC curve for this test is

$$\begin{aligned} C(p) &= P(\text{accept } H_0 \mid p) \\ &= P(X = 0 \mid p) \\ &= (1 - p)^{10}, \end{aligned}$$

again, a monotonic decreasing function of p. The size of the test then is $Q(.01) = 1 - C(.01) = 1 - (.99)^{10} = .096$; if in fact $p = .05$, the probability of a type II error is $C(.05) = (.95)^{10} = .599$, whereas if $p = .1$, the probability of a type II error is $C(.1) = (.9)^{10} = .349$ with this test. ■

Comparing tests of composite hypotheses thus involves the comparison of functions, either comparing OC functions or equivalently, power functions for competing tests; this comparison of functions is a more complex task than is involved in finding the best test of a simple H_0 versus a simple H_1. Again, in the composite case the trivial rules "always accept H_0" or "always reject H_0" could be employed, ignoring the observed sample values; you can easily verify that each of these is the best rule, for some $\theta \in \Omega$, and the worst for others, so we would not want to use either one. In the simple H_0–simple H_1 case we decided to consider all possible tests with the same value for $\alpha = P(\text{type I error})$ and, among these, called the one with the smallest value for $\beta = P(\text{type II error})$ the best test.

The same reasoning applied to the composite situation leads us to consider all tests of the same size α (α = maximum value of $Q(\theta)$ for θ values specified by H_0) and, among these, to seek the test that has the highest value for $Q(\theta)$ [equivalently, the one with the smallest value for $C(\theta) = 1 - Q(\theta)$] for all θ values specified by H_1. Such a test, if it exists, is called *uniformly most powerful*, because it maximizes the power function for all θ specified by H_1. Uniformly most powerful tests do exist for many important cases; these can frequently be found by again using the Neyman–Pearson Theorem, 8.1.1. Recall that in testing H_0: $\theta = \theta_0$ versus $H_1 : \theta = \theta_1$ (for convenience assume $\theta_1 > \theta_0$) the best test is one with critical region

$$R = \{\underset{\sim}{x}: L_X(\theta_0) \leq kL_X(\theta_1)\},$$

where k is selected to make

$$P(\text{type I error}) = P(\underset{\sim}{X} \in R \,|\, \theta = \theta_0) = \alpha;$$

also recall that we can in general simplify the implementation of the rule by finding the statistic, say, $g(\mathbf{x})$, such that

$$\{\underset{\sim}{x}: L_X(\theta_0) \leq kL_X(\theta_1)\} \quad \text{and} \quad \{\underset{\sim}{x}: g(\underset{\sim}{x}) \geq c\}$$

are equivalent (now c is chosen to fix the value for α). Suppose in testing H_0: $\theta \leq \theta_0$ versus H_1: $\theta > \theta_0$, the likelihood function $L_X(\theta)$ is such that

$$\{\underset{\sim}{x}: L_X(\theta_0) \leq kL_X(\theta_1)\}$$

is equivalent to

$$\{\underset{\sim}{x}: g(\underset{\sim}{x}) \geq c\},$$

for *every* $\theta_1 > \theta_0$, and is equivalent to

$$\{\underset{\sim}{x}: g(\underset{\sim}{x}) \leq c\}$$

for every $\theta_1 < \theta_0$ (both these inequalities may be reversed). Then it would follow that

1. The maximum value for $Q(\theta)$, for all $\theta \leq \theta_0$ (those specified by H_0) is $\alpha = Q(\theta_0)$, so the test is of size α.
2. The probability of a type II error is the smallest, for each $\theta > \theta_0$, among all tests of the same size (equivalently, the power is the greatest for all $\theta > \theta_0$).

That is, this test would be uniformly most powerful of size α. The following example should help clarify this reasoning.

EXAMPLE 8.2.3. Let $X_1, X_2, \ldots, X_n$ be a random sample of an exponential random variable; we want to test H_0: $\lambda \leq \lambda_0$ versus H_1: $\lambda > \lambda_0$, where $\lambda_0 > 0$ is some specified constant. The inequality

$$L_X(\lambda_0) \leq kL_X(\lambda_1)$$

is

$$\lambda_0^n e^{-\lambda_0 \Sigma x_i} \leq k\lambda_1^n e^{-\lambda_1 \Sigma x_i},$$

which is equivalent to

$$(\lambda_1 - \lambda_0)\sum x_i \leq \ln k + n \ln \frac{\lambda_1}{\lambda_0}.$$

Then since $\lambda_1 - \lambda_0 > 0$ for all $\lambda_1 > \lambda_0$ and $\lambda_1 - \lambda_0 < 0$ for all $\lambda_1 < \lambda_0$, this inequality is equivalent to

$$\bar{x} \leq \frac{\ln k + n \ln \lambda_1/\lambda_0}{n(\lambda_1 - \lambda_0)} = c \qquad \text{if} \qquad \lambda_1 > \lambda_0$$

$$\bar{x} \geq \frac{\ln k + n \ln \lambda_1/\lambda_0}{n(\lambda_1 - \lambda_0)} = c \qquad \text{if} \qquad \lambda_1 < \lambda_0.$$

But then the test whose critical region is specified by $R = \{\underset{\sim}{x}: \bar{x} \leq c\}$, where $P(\bar{X} \leq c | \lambda = \lambda_0) = \alpha$, is the uniformly most powerful test of size α for H_0: $\lambda \leq \lambda_0$ versus H_1: $\lambda > \lambda_0$, since it is of size α [maximum $Q(\lambda) = Q(\lambda_0)$ for $\lambda \leq \lambda_0$] and it gives the smallest probability of type II error for *every* $\lambda_1 > \lambda_0$. The test in Example 8.2.1 is in fact the best (of size .05) one can achieve in testing H_0: $\lambda \leq 1$ versus H_1: $\lambda > 1$. ■

The reasoning previously employed also gives uniformly most powerful tests for testing.

1. One-sided alternatives about p, the Bernoulli parameter.
2. One-sided alternatives about μ, the parameter of a Poisson probability law.
3. One-sided alternatives about μ, the mean of a normal probability law with σ known.
4. One-sided alternatives about σ, the standard deviation of a normal probability law with μ known.

You are asked to verify these statements in the exercises.

A more general methodology is called for in testing hypotheses about parameters of probability laws whose parameter space Ω has two or more dimensions (that is, the probability law has two or more unknown parameters). Let $\underset{\sim}{\theta}$ be the vector of parameter values of the probability law: for generality let k be the number of components in $\underset{\sim}{\theta}$. Ω then is the collection of all possible values for $\underset{\sim}{\theta}$ and is a k-dimensional space or set. One of the most frequently employed methodologies is based on the generalized likelihood ratio. Suppose we want to test a hypothesis H_0 that specifies values or ranges for some one or more parameters of the probability law, versus the alternative H_1 that simply states H_0 is false. It is instructive to think of H_0 then as saying the parameter values lie in some subset $\omega \subset \Omega$: that is, H_0: $\underset{\sim}{\theta} \in \omega$ and the alternative is H_1: $\underset{\sim}{\theta} \notin \omega$. The likelihood function is $L_X(\underset{\sim}{\theta})$. We can maximize the likelihood function by finding that $\underset{\sim}{\theta} \in \Omega$, which makes $L_X(\underset{\sim}{\theta})$ as large as possible: the values in $\underset{\sim}{\theta}$ that do this, of course, are the maximum likelihood estimates, $\hat{\underset{\sim}{\theta}}$, of the components of $\underset{\sim}{\theta}$, and $L_X(\underset{\sim}{\theta})$ is the achieved maximum of $L_X(\underset{\sim}{\theta})$ over the whole parameter space Ω. We can also maximize $L_X(\underset{\sim}{\theta})$ over the values specified by H_0: $\underset{\sim}{\theta} \in \omega$; let $L_X(\hat{\omega})$ denote the largest value for $L_X(\underset{\sim}{\theta})$, $\underset{\sim}{\theta} \in \omega$. Of course, $L_X(\hat{\omega}) \leq L_X(\hat{\underset{\sim}{\theta}})$, since in constraining the possible values for $(\underset{\sim}{\theta})$ to only those in $\omega \subset \Omega$, the achieved maximum in the restricted space must be no larger than the maximum in Ω. Thus we would always have

$$l = \frac{L_X(\hat{\omega})}{L_X(\hat{\underset{\sim}{\theta}})} \leq 1.$$

This ratio,

$$l = \frac{L_X(\hat{\omega})}{L_X(\hat{\underset{\sim}{\theta}})},$$

is called the *generalized likelihood ratio*. If in fact H_0: $\underset{\sim}{\theta} \in \omega$ is true, we would expect $L_X(\hat{\omega})$ to be roughly equal to $L_X(\hat{\underset{\sim}{\theta}})$ and the ratio l of the two maxima should be close to 1. If H_0 is not true, we might expect $L_X(\hat{\omega})$ to be considerably smaller than $L_X(\hat{\underset{\sim}{\theta}})$. The generalized likelihood ratio test criterion, for testing H_0: $\underset{\sim}{\theta} \in \omega$ versus H_1: $\underset{\sim}{\theta} \notin \omega$, uses the critical region

$$R = \left\{ \underset{\sim}{x}: l = \frac{L_X(\hat{\omega})}{L_X(\hat{\underset{\sim}{\theta}})} \leq k \right\},$$

where again $k < 1$ is chosen to make the size of the test equal to α. It can be shown (you are asked to do this in the exercises) that this generalized likelihood ratio test is identical with the Neyman–Pearson best test (Theorem 8.1.1) of simple H_0 versus simple H_1 if $\Omega = \{\underset{\sim}{\theta}_0, \underset{\sim}{\theta}_1\}$ contains only two points. The generalized likelihood ratio test criterion thus reduces to the best test in the simple versus simple case and in general gives good tests in other more complicated cases. The commonly employed tests about the parameters of a normal probability law (with the other parameter unspecified) are particular cases of this generalized likelihood ratio test criterion. The following example, and theorems, cover these cases.

EXAMPLE 8.2.4. Many light bulbs currently sold have a statement like "Average Life—750 Hours" printed on the package containing them. It would seem reasonable (for pretested bulbs) that the lifetime of bulbs made by a given manufacturer should be normal in form; the statement on the package then is referring to the mean value of that normal distribution. Suppose n bulbs are purchased and turned on until they burn out; their lifetimes could be assumed to be a random sample of size n from this normal distribution. Let us derive the likelihood ratio test of the hypothesis H_0: $\mu = 750$ versus the alternative H_1: $\mu \neq 750$. The full parameter space is $\Omega = \{(\mu, \sigma): -\infty < \mu < \infty, \sigma > 0\}$ and the restricted space specified by H_0 is $\omega = \{(750, \sigma): \sigma > 0\}$ (this is called a two-sided alternative). The generalized likelihood function for the sample is

$$L_X(\mu, \sigma^2) = \frac{\exp(-\sum(x_i - \mu)^2/2\sigma^2)}{(2\pi\sigma^2)^{n/2}}$$

To determine $L_X(\hat{\omega})$, we must assume H_0 is true and find the maximum value for L_X, that is, we want to maximize $L_X(750, \sigma^2)$ with respect to σ^2. Recalling the maximization of L_X from Section 7.1 it is easily seen that the maximizing value is

$$\hat{\sigma}^2 = \sum \frac{(x_i - 750)^2}{n}$$

and thus

$$L_X(\hat{\omega}) = e^{-n/2}\left(\frac{n}{2\pi \sum (x_i - 750)^2}\right)^{n/2}$$

To evaluate $L_X(\hat{\underline{\theta}})$, recall that the maximum likelihood estimators for μ and σ^2 are

$$\hat{\mu} = \bar{x}, \qquad \hat{\sigma}^2 = \sum (x_i - \bar{x})^2/n.$$

We find then that

$$L_X(\hat{\underline{\theta}}) = e^{-n/2}\left(\frac{n}{2\pi \sum (x_i - \bar{x})^2}\right)^{n/2}$$

thus if we are to have a probability of type I error equal to α we should reject H_0 only if

$$\frac{|\bar{x} - 750|\sqrt{n(n-1)}}{\sqrt{\sum(x_i - \bar{x})^2}} > t_{1-\alpha/2},$$

where $t_{1-\alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the t distribution with $n - 1$ degrees of freedom. ■

The preceding example is a special case of the following theorem, whose proof is left to the reader.

Theorem 8.2.1. $X_1, X_2, \ldots, X_n$ is a random sample of a normal random variable with unknown mean μ and unknown variance σ^2. Then the generalized likelihood ratio test criterion critical region R for a test of size α of H_0 versus H_1 is specified as follows for the stated H_0 and H_1.

Test	H_0	H_1	R
1.	$\mu \le \mu_0$	$\mu > \mu_0$	$\bar{x} > \mu_0 + \frac{s}{\sqrt{n}} t_{1-\alpha}$
2.	$\mu \ge \mu_0$	$\mu < \mu_0$	$\bar{x} < \mu_0 - \frac{s}{\sqrt{n}} t_{1-\alpha}$
3.	$\mu = \mu_0$	$\mu \ne \mu_0$	$\left\lvert \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right\rvert > t_{1-\alpha/2}$

■

The following example derives the generalized likelihood ratio test for the hypothesis that the variance of a normal random variable has a specified value.

EXAMPLE 8.2.5. The manufacturer of a precision scale claims that the standard deviation of measurements made by his machine will not exceed .02 mg. Assuming repeated measurements made with this machine are normally distributed, let us derive the likelihood ratio test of H_0: $\sigma \le .02$ versus the alternative H_1: $\sigma > .02$. The likelihood function of the sample is, again,

$$L_X(\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[\frac{-\sum(x_i - \mu)^2}{2\sigma^2}\right].$$

To find $L_X(\hat{\omega})$, we must maximize L_X under the assumption $\sigma \leq .02$. The maximizing value for μ is $\bar{x}$ and if

$$\hat{\sigma}^2 = \sum \frac{(x_i - \bar{x})^2}{n} \leq (.02)^2,$$

then $\hat{\sigma}$ is the maximizing value for σ; if

$$\sum \frac{(x_i - \bar{x})^2}{n} > (.02)^2,$$

then the maximizing value for σ is .02 (under the restriction $\sigma \leq .02$). Thus

$$\begin{aligned} L_X(\hat{\omega}) &= e^{-n/2} \left(\frac{n}{2\pi \sum (x_i - \bar{x})^2} \right)^{n/2} & \text{if } \sum \frac{(x_i - \bar{x})^2}{n} \leq (.02)^2 \\ &= \left(\frac{1}{2\pi(.02)^2} \right) \exp\left[-\sum \frac{(x_i - \bar{x})^2}{2(.02)^2} \right] & \text{if } \sum \frac{(x_i - \bar{x})^2}{n} > (.02)^2. \end{aligned}$$

Again, the values that maximize L_X are

$$\hat{\mu} = \bar{x}, \qquad \hat{\sigma}^2 = \sum \frac{(x_i - \bar{x})^2}{n}$$

and

$$L_X(\hat{\theta}) = \left(\frac{n}{2\pi \sum (x_i - \bar{x})^2} \right)^{n/2} e^{-n/2}.$$

Then we have

$$\begin{aligned} l = \frac{L_X(\hat{\omega})}{L_X(\hat{\underset{\sim}{\theta}})} &= 1 & \text{if } \sum \frac{(x_i - \bar{x})^2}{n} \leq (.02)^2 \\ &= \left(\frac{\sum (x_i - \bar{x})^2}{n(.02)^2} \right)^{n/2} \exp\left[\frac{n}{2} - \frac{\sum (x_i - \bar{x})^2}{2(.02)^2} \right] & \text{if } \sum \frac{(x_i - \bar{x})^2}{n} > (.02)^2. \end{aligned}$$

Figure 8.2.3 gives the graph of λ versus $\sum (x_i - \bar{x})^2/n(.02)^2$. Notice that $l < k$ is equivalent to $\sum (x_i - \bar{x})^2/n(.02)^2 > c$; if in fact H_0 is true, then $\sum (X_i - \bar{X})^2/(.02)^2$ is a χ^2 random variable with $n - 1$ degrees of freedom and

$$P(\sum (X_i - \bar{X})^2/n(.02)^2 > c) = P(\chi^2(n - 1) > cn).$$

Thus to have a probability of type I error equal to α we should choose $c = \chi^2_{1-\alpha}/n$, where $\chi^2_{1-\alpha}$ is the $100(1 - \alpha)$th percentile of the χ^2 distribution with

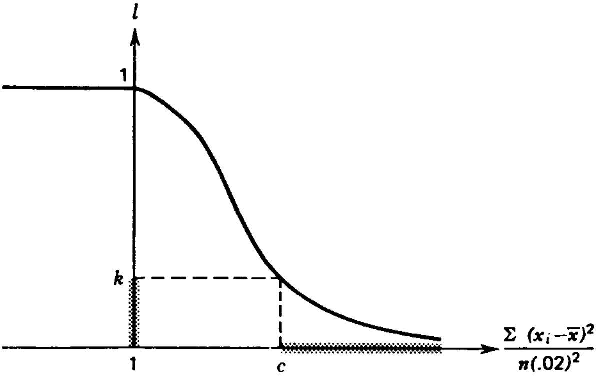

Figure 8.2.3

$n - 1$ degrees of freedom. If, for example, we used the machine $n = 10$ times to weigh the same object and we wanted to test H_0: $\sigma \leq .02$ versus H_1: $\sigma > .02$ with $\alpha = .1$, then $\chi^2_{.9} = 14.7$ (with 9 df), and we should reject H_0 if

$$\sum(x_i - \bar{x})^2 > 14.7(.02)^2 = 0.00588,$$

where $x_1, x_2, \ldots, x_{10}$ are the 10 observed measurements. ■

Theorem 8.2.2 gives the generalized likelihood ratio test criterion critical regions for testing hypotheses about the variance of a normal random variable with unknown mean.

Theorem 8.2.2. $X_1, X_2, \ldots, X_n$ is a random sample of a normal random variable X whose mean μ is unknown. Then the generalized likelihood ratio test criterion critical region R for a test of size α of H_0 versus H_1 is specified as follows for the stated H_0 and H_1.

Test	H_0	H_1	R
1.	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\sum(x_i - \bar{x})^2 > \sigma_0^2\chi^2_{1-\alpha}$
2.	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\sum(x_i - \bar{x})^2 < \sigma_0^2\chi^2_{\alpha}$
3.	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\sum(x_i - \bar{x})^2 < \sigma_0^2\chi^2_{\alpha/2}$
		and	$\sum(x_i - \bar{x})^2 > \sigma_0^2\chi^2_{1-\alpha/2}$

■

There is a direct relationship between confidence intervals for unknown parameters and tests of hypotheses about the same parameters. In fact, a $100(1-\alpha)\%$ confidence interval for an unknown parameter θ can be used to define a test of size α (one minus the confidence coefficient) for a hypothesis about θ and vice versa. To illustrate this, recall from Section 7.3 that the interval from

$$\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

to

$$\bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

is a $100(1-\alpha)\%$ confidence interval for μ, based on a random sample of size n of a normal random variable X. Suppose, in testing H_0: $\mu = \mu_0$ versus the two-sided alternative H_1: $\mu \neq \mu_0$, we use the rule: Construct the preceding $100(1-\alpha)\%$ confidence limits for μ. If μ_0 (the hypothesized value) falls in the confidence interval, accept H_0, and if it does not, reject H_0. That is, with this rule we would accept H_0: $\mu = \mu_0$ as long as

$$\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}},$$

which is easily seen to be equivalent to

$$\frac{|\bar{x}-\mu_0|}{s/\sqrt{n}} \leq t_{1-\alpha/2},$$

the same acceptance region as the generalized likelihood ratio test of size α for this hypothesis. This rule then is the same test.

Conversely, the generalized likelihood ratio test of H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$ has acceptance region (Theorem 8.2.1) defined by

$$\frac{|\bar{x}-\mu_0|}{s/\sqrt{n}} \leq t_{1-\alpha/2},$$

which is equivalent to

$$\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}.$$

So if we use this test of size α of H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$ and, based on our observed sample, define the set of μ_0 values such that we would accept H_0: $\mu = \mu_0$, this resulting set of μ_0 values is identical with the $100(1 - \alpha)\%$ confidence interval for μ, the population mean. A confidence interval for an unknown parameter can in general be translated into a test about the value for that parameter and vice versa.

EXERCISE 8.2

1. Assume the annual rainfall at a certain recording station is a normal random variable with mean μ and standard deviation 2 inches. The rainfall recorded (in inches) in each of 5 years was 18.6, 20.4, 17.3, 15.1 and 22.6. Test the hypothesis that $\mu \geq 21$ versus the alternative $\mu < 21$ with $\alpha = .1$.
2. A producer of frozen fish is being investigated by the Bureau of Fair Trades. Each package of fish that this producer markets carries the claim that it contains 12 ounces of fish; a complaint has been registered that this claim is not true. The bureau acquires 100 packages of fish marketed by this company and, letting x_i be the observed weight (in ounces) of the ith package, $i = 1, 2, \ldots, 100$, they find

$$\sum x_i = 1150, \qquad \sum x_i^2 = 13{,}249.75.$$

It would seem reasonable to assume the true weights of packages that they market are normally distributed with mean μ and variance σ^2, neither of which is known. With $\alpha = .01$, would the bureau accept or reject H_0: $\mu \geq 12$ versus H_1: $\mu < 12$, based on this sample?

3. In deciding whether a certain type of plant would be appropriate for hedges, it is of some importance that individual plants exhibit small variability in the amounts they will grow in a year (at the same age). Specifically, we might assume that the growth made by a plant of a specific type and age (for given climatic conditions) is a normal random variable with mean μ and variance σ^2. Then to decide whether the plant would be appropriate for hedges, we might like to test H_0: $\sigma^2 \geq \frac{1}{4}$ versus H_1: $\sigma^2 < \frac{1}{4}$ with $\alpha = .05$ (measurements made in feet). Suppose we record the growth of five plants of this type for 1 year and find them to be 1.9, 1.1, 2.7, 1.6 and 2.0 feet. Should we accept H_0?

*4. A manufacturer of insulated copper wire claims that his process will coat the wire so well that defects in insulation occur at a rate of no more than 1 per 100 feet. Assume ten 1000-foot rolls are examined, starting at the beginning of the roll, and the distance, X, to the first defect found is measured for each roll. If these defects occur "at random" at a constant rate of λ per foot, X is then exponential with parameter λ. Granted the 10 observed values of X were (in feet) 24, 32, 98, 584 would vote to reelect the incumbent. Letting p represent the proportion of voters in the whole electorate who would say they would reelect the incumbent, would you accept H_0: $p \geq .5$, based on this sample, with $\alpha = .1$?

13. The lubricating oil in an aircraft engine was changed on a given day; the new oil that was put into the engine contained 30 ppm iron. After 25 flight hours, $n = 11$ small samples of the oil were removed and burned on a spectrometer to estimate the current iron contamination level. The observed spectrometer readings were 34.9, 37.4, 40.1, 39.2, 34.4, 25.1, 40.7, 34.5, 30.6, 33.2, 34.0. Assuming these are the observed values of a normal random variable, would you accept H_0: $\mu = 30$ (versus H_1: $\mu \neq 30$), with $\alpha = .05$?

14. Make the same assumptions as those mentioned ~~in~~ Exercise 8.2.13 and use the data given there to test H_0: $\sigma \leq 4$ (versus H_1: $\sigma > 4$), with $\alpha = .1$.

*15. The evaluation of the power of the T-test of Theorem 8.2.1 involves the noncentral T-distribution, which we have not had the space to discuss. In testing H_0: $\mu \leq \mu_0$ versus H_1: $\mu > \mu_0$, we reject H_0 if

$$\frac{\sqrt{n}(\bar{x} - \mu_0)}{s} > t_{1-\alpha}.$$

Show why the T-distribution cannot be used to evaluate

$$P\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} > t_{1-\alpha} \middle| \mu = \mu_1\right),$$

where μ_1 is any value greater than μ_0.

8.3 Some Two-Sample Tests

In Section 7.4 we discussed sampling from two different probability laws and using the observed sample values to derive confidence limits for certain functions of the parameters of the two probability laws. In many cases it is of interest to test hypotheses about the parameters of two probability laws. For example, we might assume that a standard diet recommended for weight reduction generates normal random variables (amounts of weight lost by people using it) with mean μ_1 and variance σ^2; some new proposed diet also leads to normally distributed weight losses, say, with mean μ_2 and variance σ^2. How might we use observed sample values to test H_0: $\mu_1 \leq \mu_2$ versus H_1: $\mu_1 > \mu_2$? Or suppose the number of automobile accidents on a certain highway, per day, is assumed to be a Poisson random variable X with parameter μ_1, when the speed limit is set at 104 kilometers/hour (65 mph); if the speed limit is changed to 80 kph (50 mph), we might assume the number of accidents per day Y to be Poisson with parameter μ_2. How could we use observed values of X and Y to test H_0: $\mu_1 \le \mu_2$ versus H_1: $\mu_1 > \mu_2$? We will discuss some of the commonly used methodology for making this type of test in this section.

As you might expect after reading Section 8.2, the generalized likelihood ratio test criterion is frequently the basis for tests of hypotheses regarding the parameters of two different probability laws. We will go through this rationale in some detail for one test and then simply discuss the tests commonly used for several other cases (and whether they come from the generalized likelihood ratio test criterion).

Suppose $x_1, x_2, \ldots, x_n$ are the observed values of a random sample of a normal random variable X with mean μ_1, variance σ^2, and $y_1, y_2, \ldots, y_m$ are the observed values of an independent random sample of a normal random variable Y with mean μ_2, variance σ^2 (note the two variances are assumed equal). We want to test H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$ and will employ the generalized likelihood ratio test criterion. The vector of parameters (for the combined sample of $m + n$ values) then is $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2)$ and the parameter space is $\Omega = \{(\mu_1, \mu_2, \sigma^2)\colon -\infty < \mu_1 < \infty, -\infty < \mu_2 < \infty, \sigma^2 > 0\}$. If H_0: $\mu_1 = \mu_2$ is assumed true, the constrained parameter space is

$$\omega = \{(\mu, \mu, \sigma^2)\colon -\infty < \mu < \infty, \sigma^2 > 0\}.$$

The likelihood function is

$$\begin{aligned} L_X(\mu_1, \sigma^2)\, L_Y(\mu_2, \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{\sum(x_i - \mu_1)^2}{2\sigma^2}\right] \\ &\quad \cdot \left(\frac{1}{2\pi\sigma^2}\right)^{m/2} \exp\left[-\frac{\sum(y_j - \mu_2)^2}{2\sigma^2}\right] \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{m+n/2} \\ &\quad \cdot \exp\left[-\frac{(\sum(x_i - \mu_1)^2 + \sum(y_j - \mu_2)^2)}{2\sigma^2}\right]. \end{aligned}$$

To find the overall maximum of this likelihood function (the maximum likelihood estimates), define

$$\begin{aligned} K &= \ln L_X(\mu_1, \sigma)\, L_Y(\mu_2, \sigma) \\ &= -\frac{(m + n)}{2} \ln 2\pi - \frac{(m + n)}{2} \ln \sigma^2 - \frac{\sum(x_i - \mu_1)^2}{2\sigma^2} - \frac{\sum(y_j - \mu_2)^2}{2\sigma^2}. \end{aligned}$$

Then

$$\frac{\partial K}{\partial \mu_1} = \frac{\sum (x_i - \mu_1)}{\sigma^2}$$

$$\frac{\partial K}{\partial \mu_2} = \frac{\sum (y_j - \mu_2)}{\sigma^2}$$

$$\frac{\partial K}{\partial \sigma^2} = -\frac{m+n}{2\sigma^2} + \frac{\sum (x_i - \mu_1)^2 + \sum (y_j - \mu_2)^2}{2(\sigma^2)^2};$$

setting these partial derivatives equal to zero and solving the resulting equations simultaneously easily gives the estimates

$$\hat{\mu}_1 = \bar{x}, \qquad \hat{\mu}_2 = \bar{y}, \qquad \hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2}{m+n}.$$

Thus the unconstrained maximum value for the likelihood function is

$$L(\hat{\underline{\theta}}) = \left(\frac{m+n}{2\pi\left(\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2\right)} \right)^{(m+n)/2} e^{-(m+n)/2}.$$

With H_0 assumed true, the likelihood function becomes

$$L(\omega) = \left(\frac{1}{2\pi\sigma^2} \right)^{(m+n)/2} \exp\left[-\frac{\left(\sum (x_i - \mu)^2 + \sum (y_j - \mu)^2\right)}{2\sigma^2} \right],$$

exactly the likelihood function for a random sample of size $m + n$ of a normal random variable with mean μ and variance σ^2; thus the maximizing values are

$$\hat{\mu} = \frac{\sum x_i + \sum y_j}{m+n} = \frac{n\bar{x} + m\bar{y}}{m+n}$$

$$\hat{\sigma}^2 = \frac{\sum (x_i - \hat{\mu})^2 + \sum (y_j - \hat{\mu})^2}{m+n}$$

and

$$L(\hat{\omega}) = \left(\frac{m+n}{2\pi\left(\sum (x_i - \hat{\mu})^2 + \sum (y_j - \hat{\mu})^2\right)} \right)^{(m+n)/2} e^{-(m+n/2}.$$

The generalized likelihood ratio test criterion then is

$$l = \frac{L(\hat{\omega})}{L(\hat{\underline{\theta}})} = \left(\frac{\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2}{\sum (x_i - \hat{\mu})^2 + \sum (y_j - \hat{\mu})^2} \right)^{(m+n)/2}$$

after canceling common factors. Now recalling that

$$\hat{\mu} = \frac{n\bar{x} + m\bar{y}}{m + n},$$

we can write

$$\begin{aligned}
&\sum (x_i - \hat{\mu})^2 + \sum (y_j - \hat{\mu})^2 \\
&\quad = \sum (x_i - \bar{x} + \bar{x} - \hat{\mu})^2 + \sum (y_j - \bar{y} + \bar{y} - \hat{\mu})^2 \\
&\quad = \sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2 + n(\bar{x} - \hat{\mu})^2 + m(\bar{y} - \hat{\mu})^2 \\
&\quad = \sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2 + \frac{mn(\bar{x} - \bar{y})^2}{m + n}.
\end{aligned}$$

Dividing both numerator and denominator by $\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2$ then gives

$$l = \frac{1}{(1 + a)^{(m+n)/2}},$$

where

$$\begin{aligned}
a &= \frac{mn(\bar{x} - \bar{y})^2}{m + n} \Big/ \left(\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2\right) \\
&= \frac{t^2}{(m + n - 2)};
\end{aligned}$$

note that

$$t = (\bar{x} - \bar{y})\sqrt{\frac{mn}{m + n}} \Big/ \sqrt{\frac{\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2}{m + n - 2}}$$

is the observed value of a T random variable with $m + n - 2$ degrees of freedom if H_0: $\mu_1 = \mu_2$ is true. The critical region for the generalized likelihood ratio test criterion is defined by $l \le k$; but $l \le k$ is equivalent to $a \ge c$, which in turn is equivalent to $|t| \ge d$. Thus to have probability of type I error equal to α we should reject H_0 if

$$|\bar{x} - \bar{y}|\sqrt{\frac{mn}{m + n}} \Big/ \sqrt{\frac{\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2}{m + n - 2}} \ge t_{1-\alpha/2},$$

where the quantile is chosen from the T-distribution with $m + n - 2$ degrees of freedom. This establishes the two-sided alternative portion of Theorem 8.3.1; the one-sided alternative results also follow from the generalized likelihood ratio test criterion.

Theorem 8.3.1. Let $X_1, X_2, \ldots, X_n$ be a random sample of a normal random variable with mean μ_1, variance σ^2 and let $Y_1, Y_2, \ldots, Y_m$ be an independent random sample of a normal random variable with mean μ_2, variance σ^2. Define

$$S_p^2 = \frac{\sum (X_i - \bar{X})^2 + \sum (Y_j - \bar{Y})^2}{m + n - 2},$$

$$T = (\bar{X} - \bar{Y})\sqrt{\frac{mn}{m+n}} \bigg/ S_p.$$

Then the generalized likelihood ratio test of size α of the following hypotheses is as indicated, where the t quantiles are selected from the T distribution with $m + n - 2$ degrees of freedom.

H_0	H_1	Rejection Region
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$\lvert T \rvert \geq t_{1-\alpha/2}$
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$T > t_{1-\alpha}$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$T < t_\alpha$

■

EXAMPLE 8.3.1. Given that $n = 8$ 60-watt light bulbs of brand G provided 686, 784, 769, 848, 728, 739, 757, 743 hours of service, respectively, and that $m = 10$ 60-watt bulbs of brand W provided 762, 783, 763, 749, 806, 783, 831, 784, 790, 750 hours of service, respectively, let us use these data to illustrate the preceding theorem. Thus we assume these are two independent samples of normal random variables, both with the same variance; we want to test H_0: $\mu_1 = \mu_2$ (that the two mean lifetimes are the same) versus H_1: $\mu_1 \neq \mu_2$, with $\alpha = .05$. To make this two-sided test, then, we find $t_{.975} = 2.120$ (with $m + n - 2 = 16$ degrees of freedom), and we find from these data, $\bar{x} = 756.75$, $\bar{y} = 780.1$, $\sum(x_i - \bar{x})^2 = 15{,}555.5$, $\sum(y_j - \bar{y})^2 = 5884.9$, so the pooled (unbiased) estimate for σ^2 is

$$s_p^2 = \frac{15{,}555.5 + 5884.9}{16} = 1340.025.$$

Thus the observed T statistic is

$$t = (756.75 - 780.1)\frac{\sqrt{80/18}}{\sqrt{1340.025}}$$
$$= -1.345;$$

since $|t| = 1.345 < 2.120 = t_{.975}$, we accept H_0. ■

As mentioned at the end of Section 8.2, confidence intervals can be translated into acceptance regions for tests of hypotheses. Clearly, the test(s) described in Theorem 8.3.1 actually correspond directly to the confidence interval for $\mu_1 - \mu_2$, the difference of two normal means, discussed in Section 7.4. If you use the data given in Example 8.3.1 to evaluate a 95% two-sided confidence interval for $\mu_1 - \mu_2$, you will find the computed interval includes 0; that is, one of the possible values for $\mu_1 - \mu_2$ is 0, which is equivalent to $\mu_1 = \mu_2$. Thus the same data necessarily lead to accepting H_0: $\mu_1 = \mu_2$, with probability of type I error equal to one minus the confidence coefficient.

In Section 7.4 we also discussed $100(1 - \alpha)$% confidence intervals for σ_1^2/σ_2^2, the ratio of two normal variances. Let us convert this type of interval into a test of H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$. Again, then, let $X_1, X_2, \ldots, X_n$ be a random sample of a normal random variable X with mean μ_1 and variance σ_1^2 and let $Y_1, Y_2, \ldots, Y_m$ be an independent random sample of a normal random variable Y with mean μ_2 and variance σ_2^2. A $100(1 - \alpha)$% two-sided confidence interval for σ_2^2/σ_1^2 then has end points

$$\frac{S_Y^2}{S_X^2} F_{\alpha/2}(n-1, m-1)$$

and

$$\frac{S_Y^2}{S_X^2} F_{1-\alpha/2}(n-1, m-1).$$

The rejection region for testing H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$ then consists of those sample outcomes *not* covered by the confidence interval; that is, we reject H_0: $\sigma_1^2 = \sigma_2^2$ if the confidence interval does not include the point $\sigma_2^2/\sigma_1^2 = 1$. The confidence interval will not include 1 if the lower limit exceeds 1,

$$\frac{S_Y^2}{S_X^2} F_{\alpha/2}(n-1, m-1) > 1,$$

that is,

$$\frac{S_X^2}{S_Y^2} < F_{\alpha/2}(n-1, m-1),$$

of if the upper limit is smaller than 1,

$$\frac{S_Y^2}{S_X^2} F_{1-\alpha/2}(n-1, m-1) < 1,$$

that is

$$\frac{S_X^2}{S_Y^2} > F_{1-\alpha/2}(n-1, m-1).$$

Translating one-sided confidence limits for σ_1^2/σ_2^2 leads to tests of one-sided alternatives, summarized in Theorem 8.3.2. It can be shown that each of these tests is also given by applying the generalized likelihood ratio test criterion.

Theorem 8.3.2. Let $X_1, X_2, \dots, X_n$ be a random sample of a normal random variable X with mean μ_1, variance σ_1^2 and let $Y_1, Y_2, \dots, Y_m$ be an independent random sample of a normal random variable Y with mean μ_2, variance σ_2^2. Define

$$S_X^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2, \qquad S_Y^2 = \frac{1}{m-1}\sum(Y_j - \bar{Y})^2;$$

the generalized likelihood ratio tests (of size α) of the following hypotheses are as listed.

H_0	H_1	Rejection Region
$\sigma_1^2 \le \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$\frac{S_X^2}{S_Y^2} > F_{1-\alpha}$
$\sigma_1^2 \ge \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$\frac{S_X^2}{S_Y^2} < F_\alpha$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \ne \sigma_2^2$	$\frac{S_X^2}{S_Y^2} < F_{\alpha/2}$ or $\frac{S_X^2}{S_Y^2} > F_{1-\alpha/2}$

All quantiles used are from the F-distribution with $n-1$, $m-1$ degrees of freedom. ■

EXAMPLE 8.3.2. We will again use the data given in Example 8.3.1. To test that the two mean lifetimes were equal, we actually assumed the two variances to be equal. If this assumption were in doubt, we might first use the data to test the equality of the variances and, if we accept H_0: $\sigma_1^2 = \sigma_2^2$, then the previous test of equality of the two means can be made. Thus suppose we want to test H_0: $\sigma_1^2 = \sigma_2^2$, versus H_1: $\sigma_1^2 \ne \sigma_2^2$ with $\alpha = .02$.

Using the data from 8.3.1, we have

$$s_X^2 = \frac{15555.5}{7} = 2222.21, \qquad s_Y^2 = \frac{5884.9}{9} = 653.88$$

so

$$\frac{s_X^2}{s_Y^2} = 3.40;$$

from the F table with 7 and 9 degrees of freedom we find $F_{.01} = .149$, $F_{.99} = 5.613$, so we accept H_0: $\sigma_1^2 = \sigma_2^2$. Even though there is a sizable difference in the two estimated variances, it is not sufficiently great for us to reject equality with $\alpha = .02$. ■

If in Example 8.3.2 we had used $\alpha = .10$, we find $F_{.95} = 3.29$ and we would reject H_0: $\sigma_1^2 = \sigma_2^2$ (the probability we are wrong in rejecting H_0 is .1). But then we would not be able to use the T statistic of Example 8.3.1 in testing H_0: $\mu_1 = \mu_2$ because we have decided the two population variances are unequal. What does one do in this case to test that two normal means are equal, when their variances are not? This is a famous problem, called the Behrens–Fisher problem, one for which there is not unanimity regarding the best test to employ. One can examine the generalized likelihood ratio test for H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, but unfortunately the probability law for the ratio l in this case depends on σ_1^2/σ_2^2; if we do not know the value for σ_1^2/σ_2^2, then we cannot use the generalized likelihood ratio to find the value k that sets $P(\text{reject } H_0 \mid \mu_1 = \mu_2) = \alpha$.

One appealing approximate test, which is commonly used in this case, was first proposed by Welch, in the British journal *Biometrika* in 1937. Let us discuss his approximate procedure and apply it to the data in Example 8.3.1. Again, assume we have independent samples of sizes n and m, respectively, from two normal populations; the population means are μ_1, μ_2, respectively, and the population variances are σ_1^2, σ_2^2, not assumed equal. We want to test H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$. Welch reasoned that the difference, $\overline{X} - \overline{Y}$, has variance

$$\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$$

and that

$$S_X^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2, \qquad S_Y^2 = \frac{1}{m-1}\sum(Y_j - \overline{Y})^2,$$

respectively, give unbiased estimates for the two variances. Thus the magnitude of

$$U = \frac{(\bar{X} - \bar{Y})}{\sqrt{S_X^2/n + S_Y^2/m}}$$

would be a reasonable quantity to use in deciding whether to accept or reject H_0: $\mu_1 = \mu_2$. (If $m = n$ and $\sigma_1^2 = \sigma_2^2$, note that U has the T-distribution used in Theorem 8.3.1; with $\sigma_1^2 \neq \sigma_2^2$, U does not have a T-distribution.) Welch also then gives some further reasoning to show it is reasonable to approximate the true distribution for U by a T-distribution with degrees of freedom

$$d = (S_X^2/n + S_Y^2/m)^2 \Bigg/ \left(\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)} \right);$$

thus if we find $|u| > t_{1-\alpha/2}$, with degrees of freedom d, we reject H_0. The approximate probability of a type I error then is α, with this test. Since the degrees of freedom, d, will most likely not be an integer, we find $t_{1-\alpha/2}$ by interpolation. The procedure is illustrated in Example 8.3.3.

EXAMPLE 8.3.3. Suppose we had rejected H_0: $\sigma_1^2 = \sigma_2^2$ using the data from Example 8.3.1 and still want to test H_0: $\mu_1 = \mu_2$. We will employ Welch's procedure to do this. For the data given we have $n = 8$, $m = 10$, $\bar{x} = 756.75$, $\bar{y} = 780.1$, $s_X^2 = 2222.21$, $s_Y^2 = 653.88$; the observed value for the test statistic is

$$\begin{aligned} u &= \frac{(756.75 - 780.1)}{\sqrt{2222.21/8 + 653.88/10}} \\ &= -1.26, \end{aligned}$$

slightly smaller in magnitude than the observed T statistic of Example 8.3.1. The degrees of freedom for the approximate T-distribution is

$$\begin{aligned} d &= \left(\frac{2222.21}{8} + \frac{653.88}{10} \right)^2 \Bigg/ \left(\frac{(2222.21)^2}{64(7)} + \frac{(653.88)^2}{100(9)} \right) \\ &= 10.24; \end{aligned}$$

with 10 degrees of freedom, $t_{.975} = 2.228$, and with 11 degrees of freedom, $t_{.975} = 2.201$, so the interpolated value is

$$t_{.975} = 2.228 + .24(2.201 - 2.228) = 2.222.$$

We would still accept H_0: $\mu_1 = \mu_2$ with $\alpha = .05$ (at least approximately). If one examines the formula that determines d, the degrees of freedom, for this approximate test, it will always lie between $m + n - 2$ (the appropriate degrees of freedom with the assumption that $\sigma_1^2 = \sigma_2^2$) and the smaller of $n - 1$ and $m - 1$. Because the t quantile (with fixed area) decreases with increasing degrees of freedom, we must always accept H_0 with the Welch approximate test whenever we accept with the test given in Theorem 8.3.1 (for the same α). In cases in which we reject H_0: $\mu_1 = \mu_2$, using the test of Theorem 8.3.1, it is prudent to check whether one would still reject using the Welch approximate test whenever one is uncomfortable with the assumption $\sigma_1^2 = \sigma_2^2$. ■

In many applications one may have dependent samples of two random variables, the dependence sometimes occurring purposefully to more carefully control extraneous factors that may affect the comparison of interest. For example, suppose one were interested in investigating the effect of alcohol consumption (at a specified rate) on some reaction time (say, the length of time needed to hit a brake pedal of an automobile). One way to investigate such an effect would be to select, say, n people and measure their reaction times. Then we could also independently select a group of m people, have each person consume alcohol at the specified rate, and measure their reaction times. We would then have two independent samples and, assuming normality, could use Theorem 8.3.1 to test hypotheses of interest about μ_1 and μ_2, the two average reaction times. There are a number of facts one could criticize in this procedure. Perhaps one of the most important criticisms is that, by chance, the n people selected initially might all have naturally slow reaction times and, again by chance, the people in the group of m might all have naturally fast reaction times. This could lead us to accept H_0: $\mu_1 \geq \mu_2$, for example, simply because of the use of two independent groups. There are other criticisms that could equally be made if the preceding procedure were followed; our main interest here is simply to motivate a technique that is frequently employed in experimental work to avoid this criticism (and others as well), and to see that a reasonable model can be made incorporating dependent samples.

In the previous discussion of reaction times it probably occurred to you that a clearer investigation of the effect of alcohol consumption on reaction time could be made if we selected n people and measured their reaction time, as before. Rather than selecting a second independent group of people to consume the alcohol, we could use the same n people in the second part; each of the n consumes the alcohol and then we measure each of their reaction times a second time. This would result in n paired measurements, two reaction time measurements for n people, one made before and one made after the alcohol consumption, say, X_i = reaction time before, Y_i = reaction time after, for person i. But now it would seem reasonable that X_i and Y_i are correlated random variables, because they are reaction times for the same individual. If individual i has a natural fast reaction time, we might expect both X_i and Y_i to be above their respective averages (mean reaction times before and after). If we assume the X_i's to be normal with mean μ_1 and the Y_i's to be normal with mean μ_2, we cannot use Theorem 8.3.1 to test H_0: $\mu_1 \geq \mu_2$, for example, if we assume the two samples are correlated. Another assumption of Theorem 8.3.1 is that the two variances are equal. It is quite conceivable that the variances of reaction times before and after are not equal, again negating the use of Theorem 8.3.1.

What then might be a reasonable model, and does it lead to an easy way of testing hypotheses about the values for μ_1 and μ_2? The following model is very frequently used in such situations. We have n pairs, (X_i, Y_i), $i = 1, 2, \ldots, n$; we assume they are a random sample from a *bivariate* normal population with parameters μ_1, μ_2, σ_1^2, σ_2^2, ρ. Recall then that any linear function of X_i and Y_i is again normal; in particular, if we define $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$, the differences in the two reaction times for individual i, the D_i's are independent, normal, mean $\mu_D = \mu_1 - \mu_2$, variance $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$. Note then that $\mu_D = 0$ is equivalent to $\mu_1 = \mu_2$ and we can use the T-test of Theorem 8.2.1 to test hypotheses about $\mu_D = \mu_1 - \mu_2$. This test is called the *paired T-test* because of the natural pairings of the observations. The result is described in the following theorem.

Theorem 8.3.3. (Paired T-Test). Assume (X_i, Y_i), $i = 1, 2, \ldots, n$, is a random sample of a bivariate normal vector (X, Y) with parameters μ_1, μ_2, σ_1^2, σ_2^2, ρ; define $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$, $\mu_D = \mu_1 - \mu_2$,

$$\bar{D} = \frac{1}{n}\sum D_i, \qquad S_D^2 = \frac{1}{n-1}\sum (D_i - \bar{D})^2.$$

Then $T = (\bar{D} - \mu_D)\sqrt{n}/S_D$ has the T-distribution with $n - 1$ degrees of freedom. This distribution can be used to test the following hypotheses, with P(type I error) $= \alpha$, as follows.

H_0	H_1	Rejection Region
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$\bar{d}\sqrt{n}/s_D > t_{1-\alpha}$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$\bar{d}\sqrt{n}/s_D < t_\alpha$
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$\lvert\bar{d}\rvert\sqrt{n}/s_D > t_{1-\alpha/2}$

■

EXAMPLE 8.3.4. The following reaction times were gathered from $n = 10$ volunteers; the units used are milliseconds. For each individual the x value is the first reaction time (before consumption of beverage) and the y value is the second reaction time (after consumption) for the same individual; $d = x - y$.

Individual	1	2	3	4	5	6	7	8	9	10
$x =$	469	563	693	737	706	595	634	511	620	496
$y =$	697	814	850	933	821	788	818	761	792	763
$d =$	-228	-251	-157	-196	-115	-193	-184	-250	-172	-267

We will use these data to test H_0: $\mu_1 \geq \mu_2$ versus H_1: $\mu_1 < \mu_2$ with $\alpha =$ .01. We find $\sum d_i = -2013$, $\sum d_i^2 = 425753$, $\bar{d} = -201.3$, $s_d = 47.77$, so

$$\frac{\bar{d}\sqrt{10}}{s_d} = -13.33;$$

with 9 degrees of freedom, $t_{.01} = -2.821$, so we reject H_0. One might wonder why we chose to test H_0: $\mu_1 \geq \mu_2$. Surely it is expected that if any difference in average reaction time exists, the later reaction time should be greater than the former, meaning we would expect $\mu_1 < \mu_2$ (as we have in fact concluded). Frequently, in experimental work null hypotheses are expressed so that hopefully the data collected will lead to rejection of H_0, mainly because if we do reject H_0, the only possible error we could commit is the type I, whose maximum value is α (.01 in this case). This gives an easy way of controlling the probability that an error has been committed. Had we tested (and accepted) H_0: $\mu_1 \leq \mu_2$, the only possible error we could have committed is a type II that (as can be shown) is as small as possible, given α, for all alternatives, but we do not have so clear an indication of its maximum value for reasonable alternatives. ■

In Section 7.4 (Theorem 7.4.3) we also discussed a confidence interval for the ratio, λ_X/λ_Y, of two exponential parameters. This interval is easily converted into tests of hypotheses about the equality of λ_X and λ_Y. Assume as before $X_1, X_2, \ldots, X_n$ is a random sample of an exponential random variable with parameter λ_X, whereas $Y_1, Y_2, \ldots, Y_m$ is an independent random sample of an exponential random variable with parameter λ_Y. The $100(1-\alpha)$ % confidence limits for λ_X/λ_Y then are

$$\frac{\bar{Y}}{\bar{X}}F_{\alpha/2}, \qquad \frac{\bar{Y}}{\bar{X}}F_{1-\alpha/2},$$

where both quantiles are from the F-distribution with degrees of freedom $2n$, $2m$. If H_0: $\lambda_X = \lambda_Y$ is true, the ratio $\lambda_X/\lambda_Y = 1$, and the confidence interval for λ_X/λ_Y will not cover 1 if the lower limit exceeds 1 or if the upper limit is smaller than 1; thus the confidence interval does not include 1 if

$$\frac{\bar{Y}}{\bar{X}}F_{\alpha/2} > 1, \qquad \text{that is,} \qquad \frac{\bar{X}}{\bar{Y}} < F_{\alpha/2},$$

or if

$$\frac{\overline{Y}}{\overline{X}} F_{1-\alpha/2} < 1, \qquad \text{that is,} \qquad \frac{\overline{X}}{\overline{Y}} > F_{1-\alpha/2},$$

which defines the rejection region. It can be shown that this test is again the same as the generalized likelihood ratio test.

Theorem 8.3.4. Let $X_1, X_2, \dots, X_n$ be a random sample of an exponential random variable with parameter λ_X and let $Y_1, Y_2, \dots, Y_m$ be an independent random sample of an exponential random variable with parameter λ_Y. The generalized likelihood ratio tests of the following hypotheses are as listed; all quantiles are from the F-distribution with $2n$ and $2m$ degrees of freedom.

H_0:	H_1:	Rejection Region
$\lambda_X \le \lambda_Y$	$\lambda_X > \lambda_Y$	$\frac{\overline{X}}{\overline{Y}} < F_\alpha$
$\lambda_X \ge \lambda_Y$	$\lambda_X < \lambda_Y$	$\frac{\overline{X}}{\overline{Y}} > F_{1-\alpha}$
$\lambda_X = \lambda_Y$	$\lambda_X \ne \lambda_Y$	$\frac{\overline{X}}{\overline{Y}} < F_{\alpha/2}$ or $\frac{\overline{X}}{\overline{Y}} > F_{1-\alpha/2}$

■

EXAMPLE 8.3.5. A computer was used to generate 13 independent, exponential random variables. The first 7 values generated were 2.542, 3.508, 5.593, 5.746, .054, .243, .002 and the last 6 were 1.371, 7.655, 2.866, 2.966, 7.276, 6.144, respectively. If this generator is working as it should, these values should be just like two independent samples of sizes $n = 7$, $m = 6$, respectively, of exponential random variables with the same parameter. Let us use these values to test H_0: $\lambda_X = \lambda_Y$ versus H_1: $\lambda_X \ne \lambda_Y$ with $\alpha = .1$; the $n = 7$ values are a random sample of X, whereas the $m = 6$ values are a random sample of Y. We have $\bar{x} = \frac{17.688}{7} = 2.527$, $\bar{y} = \frac{28.278}{6} = 4.713$, so $\bar{x}/\bar{y} = .536$. With $2n = 14$, $2m = 12$ df, $F_{.05} = .395$, $F_{.95} = 2.637$ and since $\bar{x}/\bar{y}$ lies between these two values, we accept H_0; the total of the first 7 is fairly small compared to the total of the last 6, but the difference is not sufficiently great to warrant rejecting H_0: $\lambda_X = \lambda_Y$ with $\alpha = .1$ (as indeed it should not be). ■

Frequently, one may want to test hypotheses regarding the values of two Bernoulli parameters. Let us discuss a simple way of doing this for large samples from both populations; the particular test we will describe here is actually identical with a "contingency table" test that we will discuss in Chapter 10. Let $X_1, X_2, \ldots, X_n$ be a random sample of a Bernoulli random variable with parameter p_1 and let $Y_1, Y_2, \ldots, Y_m$ be an independent random sample of a Bernoulli random variable with parameter p_2. How might we test H_0: $p_1 = p_2$ versus H_1: $p_1 \neq p_2$, with a specified value for α (at least approximately)? Define $\bar{X} = (1/n) \sum X_i$, $\bar{Y} = (1/m) \sum Y_j$ and, as we know, $\bar{X}$ then is approximately normal with mean p_1, variance $p_1(1 - p_1)/n$, $\bar{Y}$ is approximately normal with mean p_2, variance $p_2(1 - p_2)/m$ and the two are independent. The difference $\bar{X} - \bar{Y}$ then is approximately normal with mean $p_1 - p_2$ and variance $p_1(1 - p_1)/n + p_2(1 - p_2)/m$; if H_0: $p_1 = p_2$ is true, the mean of this difference is 0 and the variance is $p(1 - p) \cdot (1/n + 1/m) = (m + n)\, p(1 - p)/mn$, where p is the common value for p_1 and p_2. Under the assumption $p_1 = p_2 = p$, the maximum likelihood estimate for p is

$$\hat{p} = \frac{\sum x_i + \sum y_j}{n + m} = \frac{n\bar{x} + m\bar{y}}{n + m}$$

and, still making this assumption,

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{m + n}{mn}\left(\dfrac{n\bar{X} + m\bar{Y}}{m + n}\right)\left(1 - \dfrac{n\bar{X} + m\bar{Y}}{m + n}\right)}}$$

is approximately a standard normal random variable. Thus if we reject H_0: $p_1 = p_2$ when $|Z| > z_{1-\alpha/2}$, our probability of type I error is approximately α.

This test is not equivalent to the confidence interval for $p_1 - p_2$, discussed in Section 7.4, although it is close; it is informative to see where the difference lies. The test is also not equivalent to the generalized likelihood ratio test for this hypothesis. As already mentioned, it is equivalent to a contingency table test, whose motivation will be discussed in Chapter 10.

EXAMPLE 8.3.6. Assume of $n = 50$ television sets of brand R that were sold by a given store, 20 required a service call within their one-year warranty period, whereas of $m = 40$ sets of brand Z, 12 required a service call within their one-year warranty period. Assuming the $n = 50$ R sets represent independent Bernoulli trials with p_1 = probability a service call is required, and the $m = 40$ Z sets represent independent Bernoulli trials with p_2 = probability a service call is required, should we accept H_0: $p_1 =$

p_2 (versus H_1: $p_1 \neq p_2$) with $\alpha = .1$? We can use the test just discussed to make this decision. We have $\bar{x} = \frac{20}{50} = .4$, $\bar{y} = \frac{12}{40} = .3$,

$$\frac{m+n}{mn}\left(\frac{n\bar{x}+m\bar{y}}{m+n}\right)\left(1-\frac{n\bar{x}+m\bar{y}}{m+n}\right) = \frac{90}{40(50)}\left(\frac{20+12}{90}\right)\left(1-\frac{20+12}{90}\right)$$
$$= .0103$$

so

$$z = \frac{.4-.3}{\sqrt{.0103}} = .98.$$

Since $z_{.95} = 1.64$, we accept H_0, based on these samples and conclude that the two brands appear to be equally reliable during their warranty period. ■

We have been studying the classical Neyman–Pearson system for testing hypotheses. Within this system, the basic philosophy is to consider all possible tests of the same size α, where α is to be chosen in advance; the test procedure used then is the one that minimizes the probability of type II error, or equivalently, which has maximum power, granted one can find such a test. The sample sizes involved are fixed and the test leads to either of two decisions: Accept H_0 or reject H_0.

Many statistical practitioners employ a procedure that is a slight variation on this approach, frequently called the *tests of significance* methodology (the Neyman–Pearson approach is called the methodology of *tests of hypotheses*). Let us discuss some numerical values to illustrate the tests of significance procedure. Suppose a random sample of size $n = 20$ of a normal random variable is selected and we want to test H_0: $\mu \leq 2$ versus H_1: $\mu > 2$, using the T-test of Theorem 8.3.1 with $\alpha = .05$. The rule there says then that we should compute the observed t value and reject H_0 if $t \geq 1.729$, the 95th quantile of the T-distribution with 19 degrees of freedom. Thus, if the observed t value were, say, 1.730 or 2.623 or 100 we should reject H_0, whereas if the observed t value were 1.728 or 1.7284, we should accept H_0, employing the rule as stated. The Neyman–Pearson system results in a go–no go decision, without regard to how close to rejecting (but a little short) the outcome was or by how much the critical value was exceeded. Proponents of the tests of significance approach feel that this go–no go discrete approach is wasteful of information in not giving a more continuous indication of how well the observed sample values agreed or disagreed with the stated hypothesis.

The tests of significance approach to testing a hypothesis, say, H_0: $\mu \leq 2$ versus H_1: $\mu > 2$ as before, does not set the value for α, the probability of type I error, in advance. In general, proponents will use the same test statistic (the observed t in this case) and then compute the probability of getting an observed test statistic this extreme (or more extreme) if H_0 is true; that is, for the previous case, they will observe t and then compute $\alpha' = P(T \geq t)$, the area under the T density with 19 degrees of freedom, as pictured in Figure 8.3.1. If α' is "sufficiently small," they will then proceed as if H_0 is false, otherwise, they proceed as if H_0 is true. For example, if with 19 degrees of freedom we observe $t = 1.623$ in testing H_0: $\mu \leq 2$, we find $\alpha' = .061$; this says that the probability is (no greater than) .061 of observing a T value this extreme if H_0: $\mu \leq 2$ is true. Whether H_0 is then accepted or rejected depends on the practitioner; in some applications he might reject H_0 with this α' and in others he might not.

In testing two-sided alternatives, the observed tail area must be doubled, because in that case either observed test statistics that are too large or too small would be inconsistent with H_0. For example, assume again a random sample of $n = 20$ of a normal random variable, to be used to test H_0: $\mu = 2$ versus H_1: $\mu \neq 2$. If the observed T value is $t = 1.623$, now we compute

$$\alpha' = 2P(T \geq 1.623) = .122$$

as the probability of getting an observed value this extreme (or more so); this is done because an observed $t = -1.623$ or smaller is just as extreme, if $\mu = 2$, as is the assumed $t = 1.623$ or larger.

It is, of course, easy for the tests of significance advocate to see what the tests of hypothesis advocate would decide with any fixed α. If $\alpha' \leq \alpha$, the tests of hypothesis advocate would reject H_0 and if $\alpha' > \alpha$, he or she would not. It is not possible to go the other way, though, because of the subjective element used by the tests of significance advocate. That is, knowing that H_0 was rejected with $\alpha = .10$, say, does not reveal the value for α' (except we know $\alpha' \leq .10$) and the α' user might or might not have rejected H_0 with the same observed data.

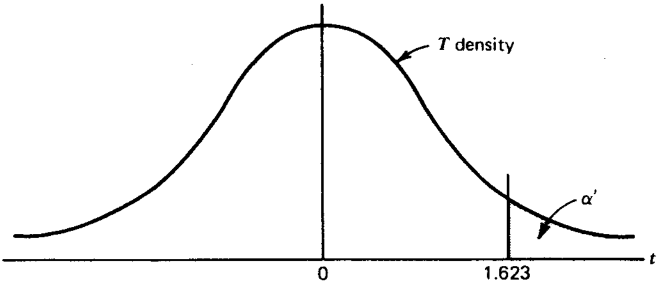

Figure 8.3.1

EXERCISE 8.3

1. To compare gas mileage capabilities of two new cars (comparable models) $n = 7$ cars of make D were driven by the same driver over the same course. The observed mileages were 22.8, 26.0, 25.6, 24.0, 25.3, 23.8, 24.5; $m = 11$ cars of make T were handled in the same way resulting in observed mileages of 19.7, 40.9, 17.2, 25.7, 40.0, 18.1, 24.5, 16.9, 26.8, 26.8, 42.8. Would you accept H_0: $\mu_D \geq \mu_T$ versus H_1: $\mu_D < \mu_T$, making the assumptions of Theorem 8.3.1?
2. Using the data of Exercise 8.3.1, would you accept the hypothesis that the variances in mileage are the same for these two makes of cars, with $\alpha = .02$? Does this result affect your conclusion in Exercise 8.3.1?
3. Ten randomly selected recent graduates of University C were selected and given an IQ test. Their scores were 120, 101, 87, 120, 107, 110, 118, 119, 112, 104. Ten recent graduates of University P were also given the same IQ test; their scores were 130, 133, 119, 123, 125, 124, 133, 120, 126, 126. Would you accept the hypothesis that the average IQ score is the same for graduates of these two universities, with $\alpha = .1$? Make the assumptions of Theorem 8.3.1.
4. Use the data of Exercise 8.3.3 to test the hypothesis that the variances in IQ scores for graduates of these two universities are equal, with $\alpha = .10$. Does the result of this test change your conclusion in Exercise 8.3.3?
5. It is assumed the number of days between earthquakes of magnitude 4.0 or more is an exponential random variable with parameter λ. On fault A the observed numbers of days between quakes of this magnitude, were 2.036, .753, .048, 5.816, 6.067, 1.449, 1.448, 1.604, respectively, for the most recent 9 earthquakes. On fault P the observed numbers of days between quakes of this magnitude were 1.972, 4.054, 2.801, 2.227, 3.826, 2.984, 1.193, 1.996, 0.982, 2.325, 3.404, respectively, for the most recent 12 earthquakes. Would you accept the hypothesis that earthquakes of this magnitude are occurring at the same rate for both faults?
6. Of 215 teenage girls admitted to a New York hospital during one month, it was found that 17 were unknowingly pregnant. Of 208 teenage girls admitted to a southern California hospital during the same month, it was found that 19 were unknowingly pregnant. Does it appear that the rate of unknowing pregnancies among teenage girls is the same in the two areas served by these hospitals?

*7. A large corporation operates two factories. It is assumed that the number of personnel accidents at each, per year, is a Poisson random variable. If there were 47 such accidents at one of the plants (in a year) and 68 accidents at the other plant (in the same year), does it appear that the two plants have the same expected number of accidents per year? (*Hint.* Develop a test based on approximate normality.)

8. The ease (or difficulty) with which one loses weight may very well be heavily dependent on the person's genetic background. To compare two different weight-loss diets, and to control the genetic contribution to how well one diet appeared versus the other, 12 identical twins were located; each person in each pair was overweight by roughly the same amount. Within each pair of twins, one of the two was selected at random and placed on diet P for 3 months; the other was placed on diet S for 3 months. At the end of this time the weight lost (in kilograms) was recorded. The results follow.

	Twin Number											
Diet	1	2	3	4	5	6	7	8	9	10	11	12
P	12.4	10.3	6.8	11.5	10.4	9.8	5.7	9.5	9.8	8.0	7.1	10.9
S	12.8	10.0	8.7	11.9	10.6	9.7	7.9	10.8	11.6	8.8	9.0	11.1

Use two different T statistics to test H_0: $\mu_P = \mu_S$ versus H_1: $\mu_P \neq \mu_S$, where μ_P, μ_S are the expected weight losses for the two diets.

9. Of 200 families watching television at a given time in New York, it was found that 45 were watching network A; of 110 families watching television at the same time in New Jersey, it was found that 32 were watching network A. Assuming these are the results of random samples, would you accept the hypothesis that network A is equally popular in both states (at this time)?

10. Evaluate the power of the test in Exercise 8.3.2 if $10\sigma_D^2 = \sigma_T^2$.

11. Using $\alpha = .1$ in Exercise 8.3.5, evaluate the power of your test if $\lambda_A = 2\lambda_P$.

*12. Use the data in Exercise 8.3.3 to test H_0: $\mu_P = \mu_C + 20$ versus H_1: $\mu_P \neq \mu_C + 20$ where μ_P, μ_C are the expected IQ scores for graduates of the two universities.

*13. Use the data in Exercise 8.3.1 to test H_0: $\sigma_T^2 = 2\sigma_D^2$.

8.4 Summary

Hypothesis: Statement about a probability law.
Simple hypothesis: Statement that uniquely identifies a probability law.
Composite hypothesis: One that is not simple.

Test of a hypothesis: Rule for deciding whether to accept or reject a hypothesis.

Critical region of test: Collection of possible observations that lead to rejection of the hypothesis.

Rejection region of test: Same as critical region.

Acceptance region of test: Complement of critical region.

Type I error: Rejecting a true hypothesis.

Type II error: Accepting a false hypothesis.

Best Neyman–Pearson test: Rule that minimizes $\beta = P(\text{type II error})$ for any fixed $\alpha = P(\text{type I error})$, when testing simple H_0 versus simple H_1 (see Theorem 8.1.1).

Equivalence of confidence intervals and tests: The values covered by the confidence interval are those such that H_0 is accepted.

Parameter space Ω: Collection of all possible values of the parameters of a probability law.

Operating characteristic function of a test: $C(\theta) = P(\text{accept } H_0 | \text{value for the parameter})$.

Power function of a test: $Q(\theta) = P(\text{reject } H_0 | \text{value for the parameter})$.

Size of a test: Maximum probability of rejecting H_0, assuming H_0 is true.

Uniformly most powerful test: One that maximizes the power function among all tests of the same size.

Generalized likelihood ratio test criterion: Test based on the value of $l = L_X(\hat{\omega})/L_X(\hat{\theta})$, where $L_X(\hat{\omega})$, $L_X(\hat{\theta})$ are the maximum values of the likelihood function, assuming H_0 is true, ignoring H_0, respectively.

Standard tests:

Mean of normal, σ^2 unknown, Theorem 8.2.1
Variance of normal, Theorem 8.2.2
Equality of two normal means, Theorem 8.3.1
Equality of two normal variances, Theorem 8.3.2
Paired T test, Theorem 8.3.3
Equality of two exponential parameters, Theorem 8.3.4.

Significance tests: Procedure in which the size is not fixed in advance; acceptance or rejection is based on the probability of observing sample as extreme, or more so, if the hypothesis is true.