Statistical Table **O,** based on the method of Tate and Klett (1959). With the help of Table **O,** which gives $(n - 1)/\chi^2_{p[n-1]}$, where p is an adjusted value of $\alpha/2$ or $1 - \alpha/2$ designed to yield the shortest unbiased confidence intervals, the computation is very simple and much faster than calculating confidence limits by means of χ^2.

Similar shortest unbiased confidence limits can be found for binomial parameters p and for the mean λ of a Poisson variable. Box 7.4 shows the simple computation by means of tables based on methods described in Crow (1956) and Crow and Gardner (1959).

7.8 Introduction to Hypothesis Testing

The most frequent application of statistics in biological research is to test a hypothesis. Statistical methods are important in biology because results of experiments are usually not clear-cut and therefore need statistical tests to support decisions between alternative hypotheses. A statistical test examines a set of sample data and, on the basis of an expected distribution of the data, leads to a decision about whether to accept the hypothesis underlying the expected distribution or reject that hypothesis and accept an alternative one. The nature of the tests varies with the data and the hypothesis, but the same general philosophy of hypothesis testing is common to all tests. Study the material in this section very carefully because it is fundamental to an understanding of every subsequent chapter in this book!

Let us consider again the sample of 17 animals of species A, 14 of which were females and 3 of which were males, which we discussed in Section 5.2. If our question is whether this litter could have come from a population with a true sex ratio of 1 : 1, we can visualize the following experiment. First we must create an infinitely sized population, half of whose members are females, the other half males. From this we can now repeatedly sample litters of 17 animals and examine how often we obtain a result of 14 females and 3 males. Since assembling an infinite population is impossible, we accomplish the same result with a finite population by sampling with replacement. This approach to hypothesis testing through randomization, almost always carried out by computer, is very common today, and, as we will see in Chapter 18, is the preferred approach in many cases.

In the example with the litter of 17 animals, however, we can simplify our task by relying on probability theory, using what we learned about the binomial distribution in Chapter 5. Assuming that the litter was sampled at random from a binomially distributed population, we concluded from Table 5.3 that if the sex ratio in the population was 1 : 1 ($p_{♀} = q_{♂} = 0.5$), the probability of obtaining a sample with 14 females and 3 males is 0.005188, making it very unlikely that such a result could be obtained by chance alone. We learned that it is conventional to include all "worse" outcomes—that is, all those that deviate even more from the outcome expected on the hypothesis $p_{♀} = q_{♂} = 0.5$. Including all

worse outcomes, the probability is 0.006363, still a very small value. This computation is based on the one-tailed test, in which we are interested only in departures from the 1 : 1 sex ratio that show a preponderance of females. If we have no preconception about the direction of the departures from expectation, we must calculate the probability of obtaining a sample as deviant as 14 females and 3 males *in either direction* from expectation—that is, the probability either of obtaining a sample of 3 females and 14 males (and all worse samples) or of obtaining 14 females and 3 males (and all worse samples). Such a test is two-tailed, and since the distribution is symmetrical, we can simply double the previously discussed probability to yield 0.012726.

What does this probability mean? Our hypothesis is that $p_{\varphi} = q_{\sigma} = 0.5$. Let us call this hypothesis H_0, the **null hypothesis,** which is the hypothesis under test. It is called the null hypothesis because it states that there is no real difference between the true value of p in the population from which we sampled and the hypothesized value of $\hat{p} = 0.5$. For instance, in the current example we believe that our sample does not exhibit a 1 : 1 sex ratio only because of sampling error. As stated earlier, this probability was computed on the assumptions that the litter was sampled at random, that the population of samples is binomially distributed, and that $\hat{p} = 0.5$. In singling out the last assumption as the null hypothesis and considering it to be improbable in view of the outcome of the experiment, we are implicitly expressing our confidence in the validity of the other two assumptions. If the litter was not sampled at random and/or the data were not binomially distributed, the probability of obtaining the observed outcome would be seriously in error. When carrying out a statistical test, we usually make several assumptions associated with the test. The investigator chooses which of these to make the null hypothesis and which others to consider valid assumptions.

If the null hypothesis $p_{\varphi} = q_{\sigma} = 0.5$ is true, then approximately 13 samples out of 1000 will be as deviant or more deviant than this one in either direction *by chance alone.* Thus, it is quite *possible* to have arrived at a sample of 14 females and 3 males or 14 males and 3 females by chance, but it is not very *probable,* since so deviant an event would occur only about 13 out of 1000 times, or 1.3% of the time. If we actually obtain such a sample, we may make one of two decisions: that the null hypothesis is true (that is, the sex ratio is 1 : 1) and that the sample obtained by us just happened to be one of those in the tail of the distribution, or that so deviant a sample is too improbable an event to justify acceptance of the null hypothesis—that is, that the hypothesis about the sex ratio being 1 : 1 is not true.

Either of these decisions may be correct, depending upon the truth of the matter. If the 1 : 1 hypothesis is correct, then the first decision (to accept the null hypothesis) will be correct. If we decide to reject the hypothesis under these circumstances, we commit an error. *The rejection of a true null hypothesis* is called a **type I error.** On the other hand, if the true sex ratio of the population is other than 1 : 1, the first decision (to accept the 1 : 1 hypothesis) is an error, a

so-called **type II error,** which is *the acceptance of a false null hypothesis.* Finally, if the 1 : 1 hypothesis is not true and we do decide to reject it, then again we make the correct decision. Thus, there are two kinds of correct decisions, accepting a true null hypothesis and rejecting a false null hypothesis, and two kinds of errors, type I, rejecting a true null hypothesis, and type II, accepting a false null hypothesis. The relationships between hypotheses and decisions can be summarized as follows:

		Null hypothesis	
		Accepted	Rejected
Null hypothesis	True	Correct decision	Type I error
	False	Type II error	Correct decision

Before we carry out a test, we have to decide what magnitude of type I error (rejection of true hypothesis) we will allow. There will always be some samples that by chance are very deviant. The most deviant of these are likely to mislead us into believing the null hypothesis to be untrue. If we permit 5% of samples to lead us into a type I error, then we will reject 5 out of 100 samples from the population, deciding that these are not samples from the given population. In the distribution under study, this means that we would reject all samples of 17 animals containing 13 of one sex and 4 of the other sex, as you can see by referring to column (3) of Table 7.2, where the expected frequencies of the outcomes on the hypothesis $p_{♀} = q_{♂} = 0.5$ are shown. This table is an extension of Table 5.3, which showed only a tail of the distribution. Actually, you would obtain a type I error slightly less than 5% if you summed relative expected frequencies for both tails starting with the class of 13 of one sex and 4 of the other. Summing the frequencies in Table 7.2, the relative expected frequency in the two tails is $2 \times 0.024{,}520{,}9 = 0.049{,}041{,}8$. In a discrete frequency distribution, such as the binomial, we cannot calculate errors of exactly 5% as we can in a continuous frequency distribution, where we can measure off exactly 5% of the area. If we decide on an approximate 1% error, we reject the hypothesis $p_{♀} = q_{♂}$ for all samples of 17 animals having 14 or more of one sex (from Table 7.2 we find that the $\hat{f}_{\text{rel}}$ in the tails sum to $2 \times 0.006{,}362{,}9 = 0.012{,}725{,}8$). Thus, the smaller the type I error we are prepared to accept, the more deviant a sample has to be for us to reject the null hypothesis H_0. Your natural inclination might be to have as little error as possible. You may decide to work with an extremely small type I error, such as 0.1% or even 0.01%, accepting the null hypothesis unless the sample is extremely deviant. The difficulty with such an approach is that although guarding against an error of the first kind, you might be falling into an error of the second kind (type II), accepting the null hypothesis when in fact it is not true and an alternative hypothesis H_1 is true. We will show later in the discussion how this comes about.

Table 7.2 RELATIVE EXPECTED FREQUENCIES FOR SAMPLES OF 17 ANIMALS UNDER TWO HYPOTHESES.

Binomial distribution.

(1) ♀♀	(2) ♂♂	(3) $H_0: p_{♀} = q_{♂} = \frac{1}{2}$ $\hat{f}_{rel}$	(4) $H_1: p_{♀} = 2q_{♂} = \frac{2}{3}$ $\hat{f}_{rel}$
17	0	0.000,007,6	0.001,015,0
16	1	0.000,129,7	0.008,627,2
15	2	0.001,037,6	0.034,508,6
14	3	0.005,188,0	0.086,271,5
13	4	0.018,158,0	0.150,975,2
12	5	0.047,210,7	0.196,267,7
11	6	0.094,421,4	0.196,267,7
10	7	0.148,376,5	0.154,210,4
9	8	0.185,470,6	0.096,381,5
8	9	0.185,470,6	0.048,190,7
7	10	0.148,376,5	0.019,276,3
6	11	0.094,421,4	0.006,133,4
5	12	0.047,210,7	0.001,533,3
4	13	0.018,158,0	0.000,294,9
3	14	0.005,188,0	0.000,042,1
2	15	0.001,037,6	0.000,004,2
1	16	0.000,129,7	0.000,000,2
0	17	0.000,007,6	0.000,000,0
Total		1.000,000,2	0.999,999,9

First let us learn some more terminology. A type I error is most frequently expressed as a probability and is symbolized by α. When expressed as a percentage it is known as the **significance level.** Thus a type I error of $\alpha = 0.05$ corresponds to a significance level of 5% for a given test. When we cut off areas proportional to α, the type I error, on a frequency distribution, the portion of the abscissa under the area that has been cut off is called the **rejection region,** or *critical region,* of a test, and the portion of the abscissa that would lead to acceptance of the null hypothesis is called the **acceptance region.** Figure 7.14A is a bar diagram showing the expected distribution of outcomes in the sex ratio example, given H_0. The dashed lines separate approximate 1% rejection regions from the 99% acceptance region.

Now let us take a closer look at the type II error, the probability of accepting the null hypothesis when it is false. If you try to evaluate the probability of a type II error, you immediately run into a problem. If the null hypothesis H_0 is false,

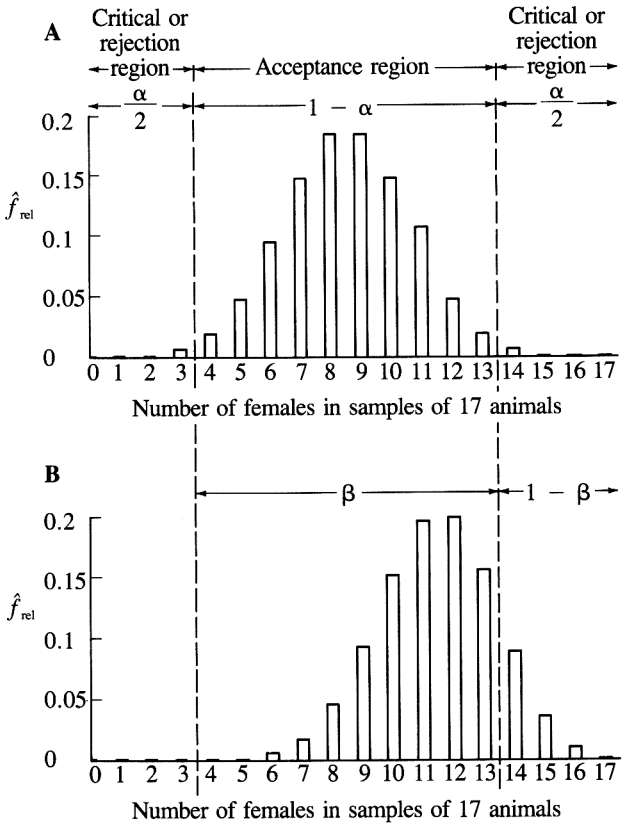

FIGURE 7.14 *Expected distributions of outcomes when sampling 17 animals from two hypothetical populations.* **A.** $H_0: p\,♀ = q\,♂ = \frac{1}{2}$. **B.** $H_1: p\,♀ = 21\,♂ = \frac{2}{3}$. *Dashed lines separate rejection, or critical, regions from the acceptance region of the distribution of* **A**. *Type I error* α *equals approximately 0.01.*

some other hypothesis H_1 must be true. But unless you can specify H_1, you are not in a position to calculate a type II error. Consider the following example. Suppose in our sex ratio case we have only two reasonable possibilities—(1) our old hypothesis $H_0: p_♀ = q_♂$, or (2) an alternative hypothesis $H_1: p_♀ = 2q_♂$, which states that the sex ratio is 2:1 in favor of females, so $p_♀ = \frac{2}{3}$ and $q_♂ = \frac{1}{3}$. We now have to calculate expected frequencies for the binomial distribution $(p_♀ + q_♂)^k = (\frac{2}{3} + \frac{1}{3})^{17}$ to find the probabilities of the outcomes under this hypothesis. These frequencies are shown graphically in Figure 7.14B and are tabulated and compared with expected frequencies of the earlier distribution in Table 7.2.

Suppose we had decided on a type I error of $\alpha \approx 0.01$, as shown in Figure 7.14A. At this significance level we would accept the H_0 for all samples of 17

having 13 or fewer animals of one sex. Approximately 99% of all samples will fall into this category. However, what if H_0 is not true and H_1 is true? Clearly, from the population represented by hypothesis H_1, we could also obtain outcomes in which one sex was represented 13 or fewer times in samples of 17. We have to calculate what proportion of the curve representing hypothesis H_1 will overlap the acceptance region of the distribution representing hypothesis H_0. In this case we find that 0.8695 of the distribution representing H_1 overlaps the acceptance region of H_0 (see Figure 7.14B). Thus, if H_1 is really true (and H_0 correspondingly false), we would erroneously accept the null hypothesis 86.95% of the time. This percentage corresponds to the proportion of samples from H_1 that fall within the limits of the acceptance regions of H_0. This proportion is called β, the type II error expressed as a proportion. In this example β is quite large. A sample of 17 animals is clearly unsatisfactory to discriminate between the two hypotheses. Although 99% of the samples under H_0 would fall in the acceptance region, 87% would do so under H_1. A single sample that falls in the acceptance region would not enable us to reach a decision between the hypotheses with a high degree of reliability. If the sample had 14 or more females, we would conclude that H_1 was correct. If it had 3 or fewer females we might conclude that neither H_0 nor H_1 was true. As H_1 approaches H_0 (as in $H_1 : p_{\varphi} = 0.55$, for example), the two distributions would overlap more and more and the magnitude of β would increase, making discrimination between the hypotheses even less likely. Conversely, if H_1 represented $p_{\varphi} = 0.9$, the distributions would be much farther apart and type II error β would be reduced. Clearly, then, the magnitude of β depends, among other things, on the parameters of the alternative hypothesis H_1 and cannot be specified without knowledge of these parameters.

When the alternative hypothesis is fixed, as in the previous example ($H_1 : p_{\varphi} = 2q_{\sigma}$), the magnitude of the type I error α we are prepared to tolerate will determine the magnitude of the type II error β. The smaller the rejection region α in the distribution under H_0, the greater the acceptance region $1 - \alpha$ in this distribution. The greater $1 - \alpha$, however, the greater its overlap with the distribution representing H_1, and hence the greater β. Convince yourself of this in Figure 7.14. By moving the dashed lines outward we are reducing the critical regions representing type I error α in diagram A. But as the dashed lines move outward, more of the distribution of H_1 in diagram B will lie in the acceptance region of the null hypothesis. Thus, by decreasing α we are increasing β and in a sense defeating our own purposes.

In most applications, scientists want to keep both of these errors small, since they do not wish to reject a null hypothesis when it is true, nor do they wish to accept it when another hypothesis is correct. We will see what steps can be taken to decrease β while holding α constant at a preset level. Note, however, that there are special applications, often nonscientific, in which one type of error is less serious than the other, and our strategy of testing or method of procedure would obviously take this into account. Thus, if you were a manufacturer producing a certain item according to specifications (which correspond to the null hypothesis

in this case), you would wish to maximize your profits and to reject as few as possible, which is equivalent to making α small. That is why in industrial statistics α is known as **producers' risk.** You might not be as concerned with samples that come from a population specified by the alternative hypothesis H_1 but that appear to conform with H_0 because they fall within its acceptance range. Such products could conceivably be marketed as conforming to specification H_0. On the other hand, as a consumer you would not mind so much a large value of α, representing a large proportion of rejects in the manufacturing process. You would, however, be greatly concerned about keeping β as small as possible, since you would not wish to accept items as conforming to H_0, which in reality were samples from the population specified by H_1 that might be of inferior quality. For this reason β is known in industrial statistics as **consumers' risk.**

Let us summarize what we have learned up to this point. When we have to carry out a statistical test, we first specify a null hypothesis H_0 and establish a significance level that corresponds to a probability of α for a type I error. In the case of the sex ratios, we defined $H_0: p_{♀} = q_{♂}$ and $\alpha \approx 0.01$. Having done this, we take a sample and test whether the sample statistic is within the acceptance region of the null hypothesis. Our sample turned out to be 14 females and 3 males. Since this sample statistic falls beyond the acceptance region, we reject the null hypothesis and conclude that this sample came from a population in which $p_{♀} \neq q_{♂}$.

If we can specify an alternative hypothesis, we can calculate the probability of type II error. In this case $H_1: p_{♀} = 2q_{♂}$ and $\beta = 0.8634$. This is the probability of accepting the null hypothesis when, in fact, the alternative hypothesis is true. In certain special situations in which the alternative hypotheses can be clearly specified, as in genetics and in this example, one might then test the previous alternative hypothesis, changing it into the null hypothesis. Thus, we might now wish to test whether the true sex ratio is $2_{♀♀} : 1_{♂}$. From Figure 7.14B it is obvious that the probability of 14 males and 3 females is very small and can be ignored. The probability of obtaining 14 or more females under the new null hypothesis is $1 - \beta$ of the old alternative hypothesis, as illustrated in Figure 7.14B. Hence this probability would be 0.1366 and, if we accept $\alpha = 0.05$, we cannot reject the new null hypothesis.

Significance levels can be varied at will by the investigator. The choices are limited, however, because for many tests cumulative probabilities of the appropriate distributions have not been tabulated. One must use published probability levels, which are commonly 0.05, 0.01, and 0.001, although several others are occasionally encountered. When a null hypothesis has been rejected at a specified level of α, we say that the sample is **significantly different** from the parametric or hypothetical population at probability $P \leq \alpha$. Generally, values of α greater than 0.05 are not considered to be **statistically significant.** A significance level of 5% ($P = 0.05$) corresponds to one type I error in 20 trials, a level of 1% ($P = 0.01$) to one error in 100 trials. Significance levels less than 1% ($P \leq 0.01$) are nearly always adjudged significant; those between 5% and 1%

may be considered significant at the discretion of the investigator. Since statistical significance has a special technical meaning (H_0 rejected at $P \leq \alpha$), we shall use the adjective *significant* only in this sense; its use in scientific papers and reports, unless such a technical meaning is clearly implied, should be discouraged. For general descriptive purposes synonyms such as important, meaningful, marked, noticeable, and others can serve to underscore differences and effects.

A brief remark on null hypotheses represented by asymmetrical probability distributions is in order here. Suppose our null hypothesis in the sex ratio case had been $H_0: p_{\varphi} = \frac{2}{3}$, as discussed above. The distribution of samples of 17 offspring from such a population is shown in Figure 7.14B. Because this distribution is clearly asymmetrical, the critical regions have to be defined independently. For a given two-tailed test we can either double the probability P of a deviation in the direction of the closer tail and compare $2P$ with α, the conventional level of significance; or we can compare P with $\alpha/2$, half the conventional level of significance. In the latter case, 0.025 is the maximal value of P conventionally considered significant.

We review now what we have learned by means of a second example, this time involving a continuous frequency distribution—the normally distributed housefly wing lengths from Table 6.1—of parametric mean $\mu = 45.5$ and variance $\sigma^2 = 15.21$. Means based on 5 items sampled from these are also distributed normally (see Figure 7.1). Assume that someone presents you with a single sample of 5 housefly wing lengths and you wish to test whether it could have come from the specified population. Your null hypothesis will be $H_0: \mu = 45.5$ or $H_0: \mu = \mu_0$, where μ is the true mean of the population from which you sampled and μ_0 stands for the hypothetical parametric mean of 45.5. Assume for the moment that we have no evidence that the variance of our sample is very much greater or smaller than the parametric variance of the housefly wing lengths. (If it were, it would be unreasonable to assume that our sample comes from the specified population. There is a critical test of the assumption about the sample variance, which we will discuss later.) The curve at the center of Figure 7.15 represents the expected distribution of means of samples of 5 housefly wing lengths from the specified population. Acceptance and rejection regions for a type I error $\alpha = 0.05$ are delimited along the abscissa. The boundaries of the rejection regions are computed as follows (remember that $t_{[\infty]}$ is equivalent to the normal distribution):

$$L_1 = \mu_0 - t_{.05[\infty]}\sigma_{\bar{Y}} = 45.5 - (1.96)(1.744) = 42.08$$

and

$$L_2 = \mu_0 + t_{.05[\infty]}\sigma_{\bar{Y}} = 45.5 + (1.96)(1.744) = 48.92$$

Thus we would consider it improbable for means less than 42.08 or greater than 48.92 to have been sampled from this population. For such sample means, we would therefore reject the null hypothesis. The test we are proposing is two-tailed because we have no a priori assumption about the possible alternatives to

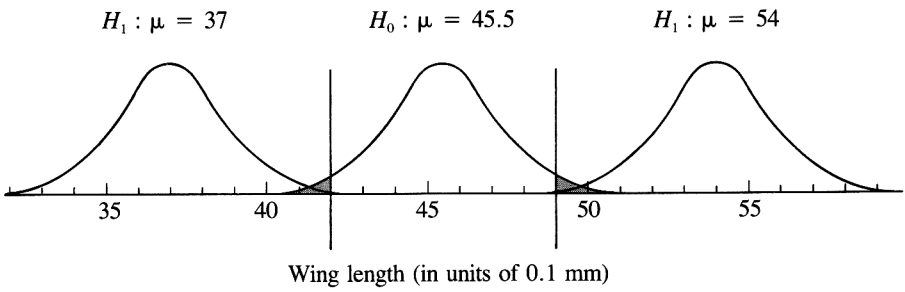

FIGURE 7.15 *Expected distribution of means of samples of 5 housefly wing lengths from normal populations specified by μ as shown above curves and $\sigma_{\bar{Y}} = 1.744$. Center curve represents null hypothesis, $H_0: \mu = 45.5$, curves at sides represent alternative hypotheses, $\mu = 37$ or $\mu = 54$. Vertical lines delimit 5% rejection regions for the null hypothesis (2½% in each tail, shaded).*

our null hypothesis. If we could assume that the true mean of the population from which the sample was taken could be only equal to or greater than 45.5, the test would be one-tailed.

Now let us examine alternative hypotheses. One alternative hypothesis might be that the true mean of the population from which our sample stems is 54.0, but that the variance is the same as before. We can express this assumption as $H_1: \mu = 54.0$ or $H_1: \mu = \mu_1$, where μ_1 stands for the alternative parametric mean 54.0. From the table of the areas of the normal curve (Table **A**) and our knowledge of the variance of the means, we can calculate the proportion of the distribution implied by H_1 that would overlap the acceptance region implied by H_0. We find that 54.0 is 5.08 measurement units from 48.92, the upper boundary of the acceptance region of H_0. This corresponds to $5.08/1.744 = 2.91\sigma_{\bar{Y}}$ units. From Statistical Table **A** we find that 0.0018 of the area will lie beyond 2.91σ at one tail of the curve. Thus under this alternative hypothesis 0.0018 of the distribution of H_1 will overlap the acceptance region of H_0. This is β, the type II error under this alternative hypothesis. Actually this is not entirely correct. Since the left tail of the H_1 distribution goes all the way to negative infinity, it will leave the acceptance region and cross over into the left-hand rejection region of H_0. However, this represents only an infinitesimal amount of the area of H_1 (the lower critical boundary of H_0, 42.08, is $6.83\sigma_{\bar{Y}}$ units from $\mu_1 = 54.0$; the area is less than 1×10^{-9}) and can be ignored.

Our alternative hypothesis H_1 specified that μ_1 is 8.5 units greater than μ_0. As we said, however, we may have no a priori reason to believe that the true mean of our sample is either greater or less than μ. Therefore we may simply assume that the true mean is 8.5 measurement units away from 45.5. In such a case we must similarly calculate β for the alternative hypothesis: $\mu_1 = \mu_0 - 8.5$. Thus the alternative hypothesis becomes $H_1: \mu = 54.0$ or 37.0, or $H_1: \mu = \mu_1$, where μ_1 represents either 54.0 or 37.0, the alternative parametric means. Since the

distributions are symmetrical, β is the same for both alternative hypotheses. Type II error for hypothesis H_1 is therefore 0.0018, regardless of which of the two alternative hypotheses is correct. If H_1 is really true, 18 out of 10,000 samples would lead to an incorrect acceptance of H_0, a very low proportion of error. These relations are shown in Figure 7.15.

You may rightly ask what reason we have to believe that the alternative parametric value for the mean is 8.5 measurement units to either side of $\mu_0 = 45.5$. Any justification for such a belief would be quite unusual. As a matter of

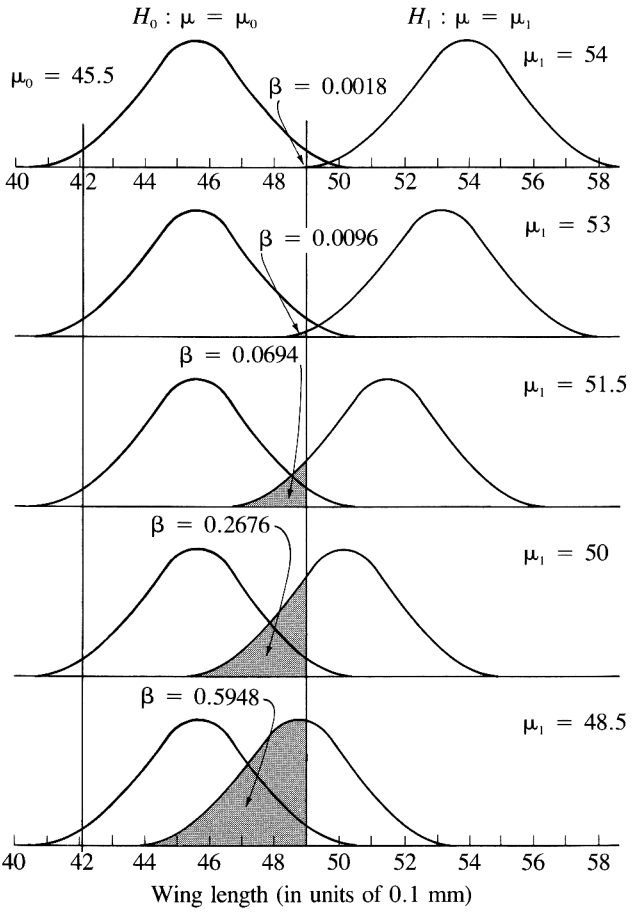

FIGURE 7.16 *Diagram to illustrate increases in type II error, β, as alternative hypothesis, H_1, approaches null hypothesis, H_0—that is, μ_1 approaches μ. Shading represents β. Vertical lines mark off 5% critical regions ($2\frac{1}{2}$% in each tail) for the null hypothesis. To simplify the graph, the alternative distributions are shown for one tail only. Data identical to those in Figure 7.15.*

fact, the true mean may just as well be 7.5 or 6.0 or any number of units to either side of μ_0. If we draw curves for $H_1: \mu = \mu_0 \pm 7.5$, we find that β has increased considerably, since the curves for H_0 and H_1 are now closer together. Thus the magnitude of β depends on how far the alternative parametric mean is from the parametric mean of the null hypothesis. As the alternative mean approaches the parametric mean, β increases to a maximum value of $1 - \alpha$, which is the area of the acceptance region under the null hypothesis. At this maximum the two distributions would be superimposed upon each other. Figure 7.16 illustrates the increase in β as μ_1 approaches μ, starting with the test illustrated in Figure 7.15. To simplify the graph, the alternative distributions are shown for one tail only. Thus we see clearly that β is not a fixed value but varies with the nature of the alternative hypothesis.

An important concept in connection with hypothesis testing is the **power** of a test. The power is $1 - \beta$, the complement of β, and is the probability of rejecting the null hypothesis when it is false and the alternative hypothesis is correct. For any given test, we would like to have the quantity $1 - \beta$ be as large as possible and the quantity β as small as possible. Since we generally cannot specify a given alternative hypothesis, we have to describe β or $1 - \beta$ for a continuum of alternative values. When $1 - \beta$ is graphed in this manner the result is called a **power curve** for the test under consideration. Figure 7.17 shows the power curve for the housefly wing length example just discussed. This figure can be compared with Figure 7.16, from which it is directly derived.

Figure 7.16 emphasizes the type II error β, and Figure 7.17 graphs the complement of this value, $1 - \beta$. Note that the power of the test falls off sharply as the alternative hypothesis approaches the null hypothesis. Common sense confirms these conclusions: We can make clear and firm decisions about whether our sample comes from a population of mean 45.5 or 60.0. The power is essentially 1. But if the alternative hypothesis is that $\mu_1 = 45.6$, differing only by 0.1 from the value assumed under the null hypothesis, deciding which of these hypotheses is true is difficult and the power will be very low.

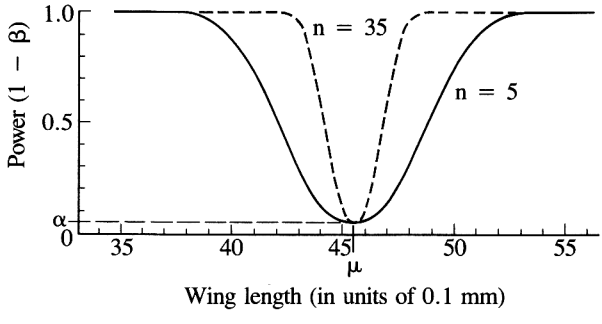

FIGURE 7.17 *Power curves for testing $H_0: \mu = 45.5$. $H_1 \mu \neq 45.5$ for $n = 5$ (as in Figures 7.15 and 7.16) and for $n = 35$.*

To improve the power of a given test (that is, decrease β) while keeping α constant for a stated null hypothesis, we must increase sample size. If, instead of sampling 5 wing lengths we had sampled 35, the distribution of means would be much narrower. Thus rejection regions for the identical type I error would now commence at 44.21 and 46.79. Although the acceptance and rejection regions would remain the same proportionately, the acceptance region would become much narrower in absolute value. Previously we could not, with confidence, reject the null hypothesis for a sample mean of 48.0. Now, when based on 35 individuals, a mean as deviant as 48.0 would occur only 15 times out of 100,000 and the hypothesis would, therefore, be rejected.

What has happened to type II error? Since the distribution curves are not as wide as before, they overlap less. If the alternative hypothesis $H_1: \mu = 54.0$ or 37.0 is true, the probability that the null hypothesis could be accepted by mistake (type II error) is infinitesimally small. If we let μ_1 approach μ_0, β will increase, of course, but it will always be smaller than the corresponding value for sample size $n = 5$. This comparison is shown in Figure 7.17, where the power for the test with $n = 35$ is much higher than that for $n = 5$. If we were to increase our sample size to 100 or 1000, the power would be increased still further. Thus we reach an important conclusion: If a given test is not sensitive enough, we can increase its sensitivity (= power) by increasing sample size.

There is yet another way of increasing the power of a test. If we cannot increase sample size, we may increase the power by changing the nature of the test. Different statistical techniques testing roughly the same hypothesis may differ substantially in both the magnitude and the slopes of their power curves. Tests that maintain higher power levels over substantial ranges of alternative hypotheses are clearly to be preferred. The popularity of the nonparametric tests mentioned in several places in this book has grown not only because of their computational simplicity but in many cases also because their power curves are less affected by failure of assumptions than are those of the parametric methods. Nonparametric tests, however, have lower overall power than do parametric ones, when all the assumptions of the parametric test are met.

While on the subject of different tests we should discuss briefly the **size** of a test. When we follow the protocols of a given statistical test and decide on a specified type I error α, the correctness of that value depends on how sensitive the test is to departures from its assumptions in the sample analyzed. The intended type I error rate may be only nominal, and the actual error rate—termed, somewhat infelicitously, the size of the test—may be greater or smaller. In the former case we speak of a **liberal test** (i.e., we reject the null hypothesis more often than we should), the latter case represents a **conservative test** (i.e., we reject the null hypothesis less often).

Let us look briefly at a one-tailed test. The null hypothesis is $H_0: \mu_0 = 45.5$ as before. The alternative hypothesis, however, assumes that we have reason to believe that the parametric mean of the population from which our sample has been taken cannot be less than $\mu_0 = 45.5$. If it is different from that value, it can be only greater than 45.5. We might have two grounds for such an hypothesis.

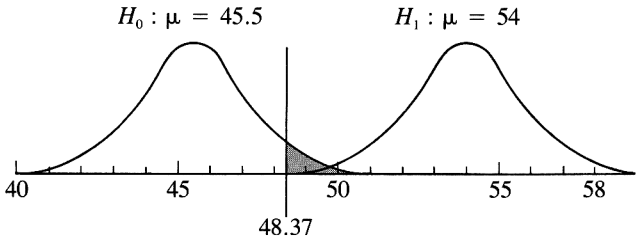

FIGURE 7.18 *One-tailed significance test for the distribution of Figure 7.15. The vertical line now cuts off 5% rejection region from one tail of the distribution (shaded area).*

First, we might have a biological reason for such a belief. For example, our parametric flies might be a dwarf population, making any other population from which our sample could have come necessarily bigger. A second reason might be that we are interested in only one direction of difference. For example, we may be testing the effect of a chemical in the larval food intended to increase the size of the sample of flies. Therefore, we would expect that $\mu_1 \geq \mu_0$, and we are not interested in testing for any μ_1 that is less than μ_0 because such an effect is the opposite of what we anticipate. Similarly, if we are investigating the effect of a certain drug as a cure for cancer, we might wish to compare the untreated population that has a mean fatality rate θ (from cancer) with the treated population, whose rate is θ_1. Our alternative hypotheses would be $H_1 : \theta_1 < \theta$. That is, we would not be interested in any θ_1 that is greater than θ because if our drug increases mortality from cancer, it certainly is not much of a prospect for a cure.

When such a one-tailed test is performed, the rejection region along the abscissa is only under one tail of the curve representing the null hypothesis. Thus, for our housefly data (distribution of means of sample size $n = 5$) the rejection region will be in one tail of the curve only and for a 5% type I error will appear as shown in Figure 7.18. We compute the critical boundary as $45.5 + (1.645)(1.744) = 48.37$. (The 1.645 is $t_{.10[\infty]}$, which corresponds to the 5% value for a one-tailed test.) Compare this rejection region, which rejects the null hypothesis for all means greater than 48.37, with the two rejection regions in Figure 7.16, which reject the null hypothesis for means lower than 42.08 and greater than 48.92. In Figure 7.18 the alternative hypothesis is considered for one tail of the distribution only, and the power curve of the test is not symmetrical but is drawn with respect to only one side of the distribution.

7.9 TESTS OF SIMPLE HYPOTHESES USING THE NORMAL AND *t*-DISTRIBUTIONS

We will now apply our new knowledge of hypothesis testing to some simple examples involving the normal and *t*-distributions.