5 Statistical Inference: Estimation and Hypothesis Testing

Parameter estimates obtained from samples are usually meant to be used to estimate the true population parameters. The sample mean and variance are typical estimators or predictors of the true mean and variance, and are often called "point" estimates. In addition, an interval that is apt to contain the true parameter often accompanies and complements the point estimate. These intervals, known as *confidence intervals*, can be constructed with a known a priori probability of bracketing the true parameters. Confidence intervals play an important role in the evaluation of drugs and drug products.

The question of *statistical significance* pervades much of the statistics commonly used in pharmaceutical and clinical studies. Advertising, competitive claims, and submissions of supporting data for drug efficacy to the FDA usually require evidence of superiority, effectiveness, and/or safety based on the traditional use of statistical hypothesis testing. This is the technique that leads to the familiar statement, "The difference is statistically significant" (at the 5% level or less, for example), words that open many regulatory doors. Many scientists and statisticians feel that too much is made of testing for statistical significance, and that decisions based on such statistical tests are often not appropriate. However, testing for statistical significance is one of the backbones of standard statistical methodology and the properties and applications of such tests are well understood and familiar in many experimental situations. This aspect of statistics is not only important to the pharmaceutical scientist in terms of applications to data analysis and interpretation, but is also critical to an understanding of the statistical process. Since much of the material following this chapter is based largely on a comprehension of the principles of hypothesis testing, the reader is urged to make special efforts to understand the material presented in this chapter.

5.1 STATISTICAL ESTIMATION (CONFIDENCE INTERVALS)

We will introduce the concept of statistical estimation and confidence intervals before beginning the discussion of hypothesis testing. Scientific experimentation may be divided into two classes: (a) experiments designed to estimate some parameter or property of a system, and (b) comparative experiments, where two or more treatments or experimental conditions are to be compared. The former type of experiment is concerned with estimation and the latter is concerned with hypothesis testing.

The term *estimation* in statistics has a meaning much like its meaning in ordinary usage. A population parameter is estimated based on the properties of a sample from the population. We have discussed the unbiased nature of the sample estimates of the true mean and variance, designated as $\overline{X}$ and S^2 (sects. 1.4 and 1.5). These sample statistics estimate the population parameters and are considered to be the best estimates of these parameters from several points of view.* However, the reader should understand that statistical conclusions are couched in terms of probability. Statistical conclusions are not invariant as may be the case with results of mathematical proofs. Without having observed the entire population, one can never be sure that the sample closely reflects the population. In fact, as we have previously emphasized, sample statistics such as the mean and variance are rarely equal to the population parameters.

* These "point" estimates are unbiased, consistent, minimum variance estimates. Among unbiased estimators, these have minimum variance, and approach the true value with high probability as the sample size gets very large.

Nevertheless, the sample statistics (e.g., the mean and variance) are the best estimates we have of the true parameters. Thus, having calculated $\overline{X}$ and S^2 for potencies of 20 tablets from a batch, one may very well inquire about the true average potency of the batch. If the mean potency of the *20 tablets is 49.8 mg*, the *best estimate* of the true batch mean is *49.8 mg*. This is known as the point estimate. Although we may be almost certain that the true batch mean is not exactly 49.8 mg, there is no reason, unless other information is available, to estimate the mean to be a value different from 49.8 mg.

The discussion above raises the question of the reliability of the sample statistic as an estimate of the true parameter. Perhaps one should hesitate in reporting that the true batch mean is 49.8 mg based on data from only 20 tablets. One might question the reliability of such an estimate. The director of quality control might inquire: "How close do you think the true mean is to 49.8 mg?" Thus, it is a good policy when reporting an estimate such as a mean to include some statement as to the reliability of the estimate. Does the 49.8-mg estimate mean that the true mean potency could be as high as 60 mg, or is there a high probability that the true mean is not more than 52 mg? This question can be answered by use of a *confidence interval.* A confidence interval is an interval within which we believe the true mean lies. We can say, for example, that the true batch mean potency is between 47.8 and 51.8 mg with 95% probability. The width of the interval depends on the properties of the population, the sample estimates of the parameters, and the degree of certainty desired (the probability statement).

Since most of the problems that we will encounter are concerned with the normal distribution, particularly sampling of means, we are most interested in confidence intervals for means. If the distribution of means is normal and σ is known, an interval with confidence coefficient, P (probability), can be computed using a table of the cumulative standard normal distribution, Table IV.2. A two-sided confidence interval, symmetric about the observed mean, is calculated as follows:

$$P\ \%\ \text{confidence interval} = \overline{X} \pm \frac{Z_p \sigma}{\sqrt{N}} \tag{5.1}$$

where $\overline{X}$ is the observed sample mean, N the sample size, σ the population standard deviation, and Z_p the normal deviate corresponding to the $(P+1)/2$ percentile of the cumulative standard normal distribution (Table IV.2).

For the most commonly used 95% confidence interval, $Z = 1.96$, corresponding to $(0.95 + 1)/2 = 0.975$ of the area in the cumulative standard normal distribution. Other common confidence coefficients are 90% and 99%, having values of Z equal to 1.65 and 2.58, respectively. The probability statement, for example, 90%, 95%, 99%, depends on the context. Therefore, one cannot say that one probability is "better" than another. For example, in bioequivalence studies, a 90% confidence interval is most appropriate (see chap. 11). Inspection of Table IV.2 shows that the area in the tails of a normal curve between $\pm$ 1.65, $\pm$ 1.96, and $\pm$2.58 standard deviations from the mean is 90%, 95%, and 99%, respectively. This is illustrated in Figure 5.1 (see also Table 3.4).

Before presenting examples of the computation and use of confidence intervals, the reader should take time to understand the concept of a confidence interval. The confidence interval changes depending on the sample chosen because, although σ[†] and N remain the same, $\overline{X}$ varies from sample to sample. A confidence interval using the mean from any given sample may or may not contain the true mean. Without knowledge of the true mean, we cannot say whether or not any given interval contains the true mean. However, it can be proven that when intervals are constructed according to Eq. (5.1), P% (e.g., 95%) of such intervals will contain the true mean. Figure 5.2 shows how means of size N, taken from the same population, generate confidence intervals. Think of this as means of size 20, each mean generating a confidence interval [Eq. (5.1)]. For a 95% confidence interval, 19 of 20 such intervals will cover the true mean, μ, on the average. Any single interval has a 95% chance of covering the true mean, a

[†] σ is assumed to be known in this example.

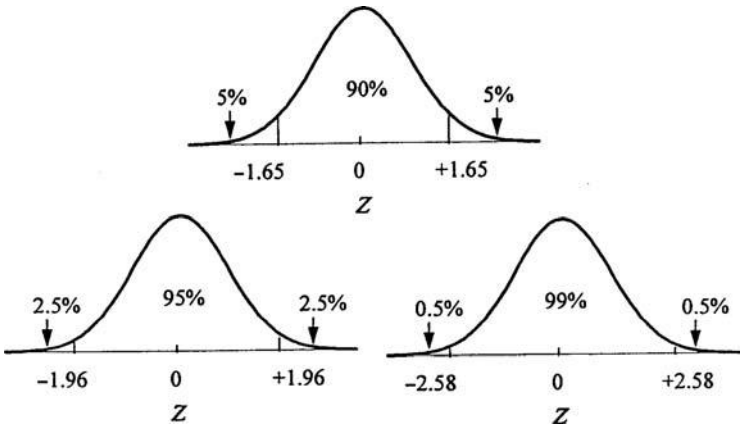


Figure 5.1 Areas in the tails of a standard normal curve.

priori. Of course, one would not usually take many means in an attempt to verify this concept, which can be proved theoretically. Under usual circumstances, only a single mean is observed and a confidence interval computed. This interval may not cover the true mean, but we know that 19 of 20 such intervals will cover the true mean.

Looking at the confidence interval from another point of view, suppose that a mean of 49.8 mg was observed for a sample size of 20 with $\sigma/\sqrt{N}$, $(\sigma_{\overline{x}})$ equal to 2. According to Eq. (5.1), the 95% confidence interval for the true mean is $49.8 \pm 1.96(2) = 45.9$ to 53.7 mg. Figure 5.3 shows that if the true mean were outside the range 45.9 to 53.7, the observation of the sample mean, 49.8 mg, would be very unlikely. The dashed curve in the figure represents the distribution of means of size 20 with a true mean of 54.7 and $\sigma_{\overline{x}} = 2$. In this example, the true mean is outside the 95% confidence interval, and the probability of observing a mean from this distribution as small as 49.8 mg or less is less than 1% (see Exercise Problem 1). Therefore, one could conclude that the true mean is probably not as great as 54.7 mg based on the observation of a mean of 49.8 mg from a sample of 20 tablets.

5.1.1 Confidence Intervals Using the *t* Distribution

In most situations in which confidence intervals are computed, σ, the true standard deviation, is unknown, but is estimated from the sample data. A confidence interval can still be computed based on the sample standard deviation, S. However, the interval based on the sample standard deviation will tend to be wider than that computed with a known standard deviation. This is reasonable because if the standard deviation is not known, one has less knowledge of the true distribution and consequently less assurance of the location of the mean.

The computation of the confidence interval in cases where the standard deviation is estimated from sample data is similar to that shown in Eq. (5.1) except that a value of t is substituted for the Z value

$$P\% \text{ confidence interval} = \overline{X} \pm \frac{tS}{\sqrt{N}}. \tag{5.2}$$

Values of t are obtained from the cumulative t table, Table IV.4, corresponding to a P% confidence interval.

The appropriate value of t depends on degrees of freedom (d.f.), a concept that we encountered in section 1.5.2. When constructing confidence intervals for means, the d.f. are equal to $N - 1$, where N is the sample size. For samples of size 20, d.f. $= 19$ and the appropriate values of t for 90%, 95%, or 99% confidence intervals are 1.73, 2.09, and 2.86, respectively. Examination of the t table shows that the values of t decrease with increasing d.f., and approach the corresponding Z values (from the standard normal curve) when the d.f. are large. This is expected, because when d.f. $= \infty$, the standard deviation is known and the t distribution coincides with

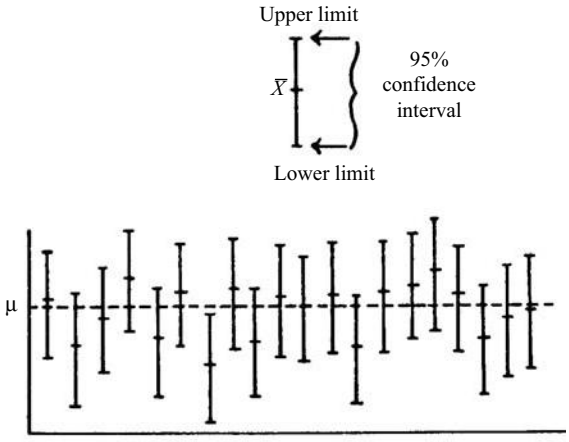


Figure 5.2 Concept of the confidence interval.

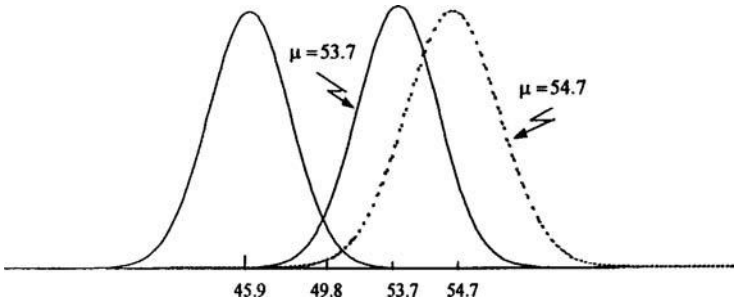


Figure 5.3 This figure shows that a mean of 49.8 is unlikely to be observed if the true mean is 54.7 (confidence interval = 45.9–53.7).

the standard normal distribution. We will talk more of the t distribution later in this chapter (see also sect. 3.5).

5.1.2 Examples of Construction of Confidence Intervals

Example 1: Confidence interval when σ *is unknown and estimated from the sample*. The labeled potency of a tablet dosage form is 100 mg. Ten individual tablets are assayed according to a quality control specification. The 10 assay results shown in Table 5.1 are assumed to be sampled from a normal distribution. The sample mean is 103.0 mg and the standard deviation is 2.22. A 95% confidence interval for the true batch mean [Eq. (5.1)] is

$$103 \pm 2.26\left(\frac{2.22}{\sqrt{10}}\right) = 101.41 \text{ to } 104.59.$$

Table 5.1 Assay Results for 10 Randomly Selected Tablets (mg)

101.8	104.5
102.6	100.7
99.8	106.3
104.9	100.6
103.8	105.0
$\bar{X} = 103.0$	$S = 2.22$

Note that the t value is 2.26. This is the value of t with 9 d.f. ($N = 10$) for a 95% confidence interval taken from Table IV.4.

Example 2: Confidence interval when σ is known. Suppose that the standard deviation were known to be equal to 2.0. The 95% confidence interval for the mean is [Eq. (5.1)]

$$\overline{X} \pm \frac{1.96\sigma}{\sqrt{N}} = 103.0 \pm \frac{1.96(2.0)}{\sqrt{10}} = 101.76 \text{ to } 104.24.$$

The value 1.96 is obtained from Table IV.2 ($Z = 1.96$ for a two-sided symmetrical confidence interval) or from Table IV.4 for t with ∞ d.f.

Two questions arise from this example.

1. How can we know the s.d. of a batch of tablets without assaying every tablet?
2. Why is the s.d. used in Example 2 different from that in Example 1?

Although it would be foolhardy to assay each tablet in a batch (particularly if the assay were destructive, that is, the sample is destroyed during the assay process), the variance of a "stable" process can often be precisely estimated by averaging or pooling the variance over many batches (see also sect. 12.2 and App. I). The standard deviation obtained from this pooling is based on a large number of assays and will become very stable as long as the tableting process does not change. The pooled standard deviation can be assumed to be equal to or close to the true standard deviation (Fig. 5.4).

The answer to the second question has actually been answered in the previous paragraph. The variance of any single sample of 10 tablets will not be identical to the true variance, 2^2 or 4 in the example above. If the average variance over many batches can be considered equal to or very close to the true variance, the pooled variance is a better estimate of the variance than that obtained from 10 tablets. This presupposes that the variance does not change from batch to batch. Under these conditions, use of the pooled variance rather than the individual sample variance will result in a narrower confidence interval, on the average.

Example 3: Confidence Interval for a Proportion.(a) In a preclinical study, 100 untreated (control) animals were observed for the presence of liver disease. After six months, 25 of these animals were found to have the disease. We wish to compute a 95% confidence interval for

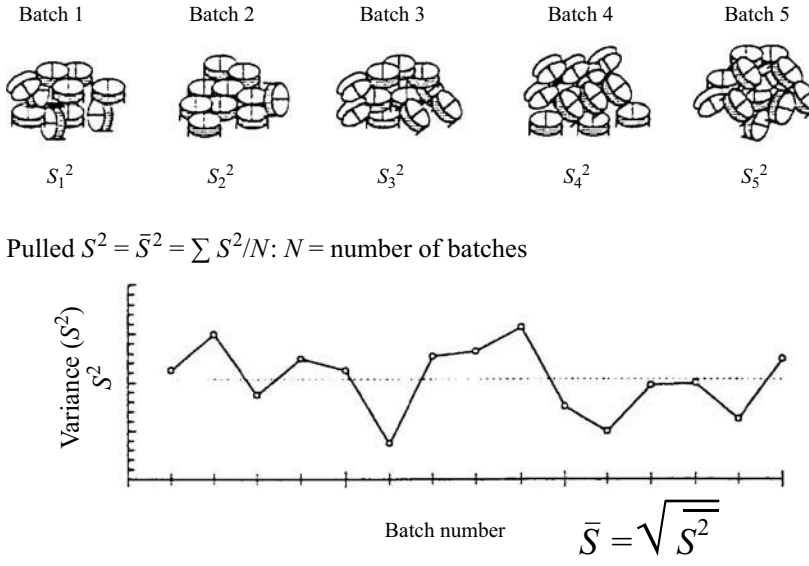


Figure 5.4 Pooling variances over batches, a good estimate of the true variance of a stable process (same sample size per batch).

the true proportion of animals who would have this disease if untreated (after six months). A confidence interval for a proportion has the same form as that for a mean. Assuming that the normal approximation to the binomial is appropriate, the confidence interval is approximately

$$\hat{p} \pm Z\sqrt{\frac{\hat{p}\hat{q}}{N}}, \tag{5.3}$$

where $\hat{p}$ is the observed proportion, $\hat{q} = 1 - \hat{p}$, Z the appropriate cutoff point from the normal distribution (Table IV.2), and N the sample size.

In the present example, a 95% confidence interval is

$$0.25 \pm 1.96\sqrt{\frac{(0.25)(0.75)}{100}} = 0.165 \text{ to } 0.335.$$

The true proportion is probably between 16.5% and 33.5%.‡ Notice that the mean is equal to the observed proportion and that the normal approximation to the binomial distribution makes use of the Z value of 1.96 for the 95% confidence interval from the cumulative normal distribution. The standard deviation is computed from Eq. (3.11), $\sigma = \sqrt{\hat{p}\hat{q}/N}$.

A 99% confidence interval for the true proportion is

$$0.25 \pm 2.58\sqrt{\frac{(0.25)(0.75)}{100}} = 0.138 \text{ to } 0.362.$$

Note that the 99% confidence interval is wider than the 95% interval. The greater the confidence, the wider is the interval. To be 99% "sure" that the true mean is contained in the interval, the confidence interval must be wider than that which has a 95% probability of containing the true mean.

(b) To obtain a confidence interval for the true *number of animals* with liver disease when a sample of 100 shows 25 with liver disease, we use the standard deviation according to Eq. (3.12), $\sigma = \sqrt{N\hat{p}\hat{q}}$. A 95% confidence interval for the true *number* of diseased animals (where the observed number is $N\hat{p} = 25$) is

$$\begin{aligned} N\hat{p} \pm 1.96\sqrt{N\hat{p}\hat{q}} &= 25 \pm 1.96\sqrt{(100)(0.25)(0.75)} \\ &= 16.5 \text{ to } 33.5. \end{aligned}$$

This answer is exactly equivalent to that obtained using proportions, in part (a) ($16.5/100 = 0.165$ and $33.5/100 = 0.335$). Further examples of symmetric confidence intervals are presented in conjunction with various statistical tests in the remaining sections of this chapter. In particular, confidence intervals for the true difference of two means or two proportions are given in sections 5.2.2, 5.2.3, and 5.2.6.

An interesting, special confidence interval that is useful for proportions is the case where 0 successes or failures are observed in N trials. For example, this situation arises in data from quality control and clinical trials. When inspecting individual items for sterility, we may observe zero defects in 1000 items inspected. In a clinical trial, we may observe no side effects of a particular kind in 200 patients. In these cases, it is of interest to put an upper bound on the proportion of failures, where failure in the above examples is the observation of a nonsterile item or a particular side effect. This can be calculated using a confidence interval. In these situations, we observe 100% successes and 0% failures, and the lower confidence interval is 0%. (We cannot have less than 0 failures.) We will put an upper limit on the true proportion of failures, equal to a lower limit on the proportion of successes. This may be thought of as a

‡ Both $N\hat{p}$ and $N\hat{q}$ should be equal to or greater than 5 when using the normal approximation to the binomial (sect. 3.4.3).

one-sided confidence interval. To compute the probability of 0 failures in N trials, we can apply the binomial formula (from chap. 3)

$$\text{Probability of } X \text{ successes in } N \text{ trials} = \binom{N}{X} P^x q^{N-X}. \tag{3.9}$$

When $X = N$ (i.e., all N trials are successes), the probability of $X = N$ successes is P^N.

To obtain a lower limit for the proportion of successes based on a 95% confidence interval, we compute the value of p that results in a probability of 0.05, when we have all the observations successful. This computation uses the following formula:

$0.05 = p^N$.
A 95% confidence interval for the true proportion of failure is $1 - p$.

Suppose that inspection of 1000 ampoules in a batch of 30,000 shows that all items are sterile. With 95% confidence what is the upper limit of potential nonsterile ampoules in the batch.

$0.05 = p^{1000}$. The log of p can be calculated as $\log(0.05)/1000 = -0.002996$. $p =$ the antilog of $-0.002996 = 0.997$. The upper limit for the proportion of failures is $1 - 0.997 = 0.003$ or 3 in a thousand. We conclude that with 95% probability, there are no more than 3 failures in 1000 items. We see that it is impossible to guarantee 100% successes without inspecting each item (see also sampling, chap. 4). Certainly, this is not possible if the sampling is destructive. Of course, the intensity of sampling is dependent on cost and the potential risks to the consumer (patients) if failures exist in the batch.

Another useful application is the Negative Binomial, Time to Failure, described in chapter 3.

5.1.3 Asymmetric Confidence Intervals

5.1.3.1 One-Sided Confidence Intervals

In most situations, a two-sided confidence interval symmetric about the observed mean seems most appropriate. This is the shortest interval given a fixed probability. However, there are examples where a one-sided confidence interval can be more useful. Consider the case of a clinical study in which 18 of 500 patients treated with a marketed drug report headaches as a side effect. Suppose that we are only concerned with an "upper limit" on the proportion of drug-related headaches to be expected in the population of users of the drug. In this example, when constructing a 95% interval, we use a Z (or t) value that cuts off 5% of the area in the upper tail of the distribution, rather than the 2.5% in each tail excluded in a symmetric interval. Using the normal approximation to the binomial, the upper limit is

$$p + Z\sqrt{\frac{pq}{N}} = \frac{18}{500} + 1.65\sqrt{\frac{(0.036)(0.964)}{500}}$$

$$= 0.036 + 0.014 = 0.050.$$

Based on the one-sided 95% confidence interval, we conclude that the true proportion of headaches among drug users is probably not greater than 5%. Note that we make no statement about the lower limit, which must be greater than 0. Another application of a one-sided confidence interval is presented in section 7.5, as applied to the analysis of stability data. If a one-sided confidence interval is to be used for regulatory decisions or other "official" applications, the rationale for using a one-sided rather than a two-sided interval should be clearly explained prior to the experiment (see also sect. 5.1.3 explaining one-sided confidence intervals).

5.1.3.2 Other Asymmetric Confidence Intervals

In general, many P% confidence intervals can be constructed by suitably allocating $(1 - P)$% of the area to the lower and upper tails of the normal distribution. For example, a 95% confidence

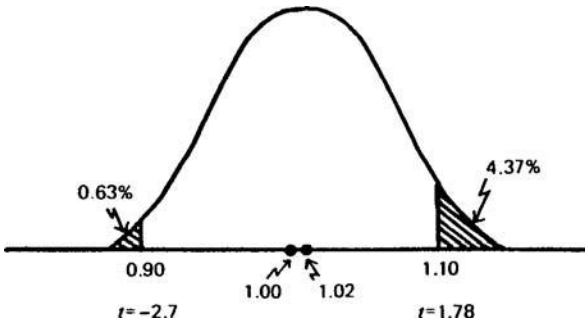


Figure 5.5 A 95% asymmetric confidence interval with $\overline{X} = 1.02$, s.d. $= 0.2$, and $N = 20$.

interval may be constructed by placing 1% of the area in the lower tail and 4% in the upper tail. This is not a common procedure and a good reason should exist before one decides to make such an allocation. Westlake [1,2] has proposed such an interval for the construction of confidence intervals in bioequivalence studies. In these studies, a ratio of some property (such as maximum serum concentration) of two products is compared. Westlake argues that an interval symmetric about the ratio 1.0 is more useful than one symmetric about the observed sample mean. The interval often has the great majority of the area in either the lower or upper tail, depending on the observed ratio. For a ratio greater than 1.0, most of the area will be in the upper tail and vice versa. Figure 5.5 illustrates this concept with a hypothetical example for products with an average ratio of 1.02. If the standard deviation is unknown and is estimated as 0.2 with 19 d.f. ($N = 20$), a 95% symmetric interval would be estimated as

$$1.02 \pm \frac{(2.1)(0.2)}{\sqrt{20}} = 1.02 \pm 0.094 = 0.926 \text{ to } 1.114.$$

To construct the Westlake interval, a symmetric interval about 1.0, detailed tables of the t distribution are needed [1]. In this example, t values of approximately 1.78 and -2.70 will cutoff 4.3% of the area in the upper tail and 0.7% in the lower tail, respectively. This results in an upper limit of $1.02 + 0.08 = 1.10$ and a lower limit of $1.02 - 0.12 = 0.90$, symmetric about 1.0 (1.0 ± 0.1).

Examples of confidence intervals for bioequivalence testing are given in chapters 11 and 15.

The remainder of this chapter will be concerned primarily with testing hypotheses, categorized as follows:

1. Comparison of the mean of a single sample (group) to some known or standard mean [single-sample (group) tests].
2. Comparison of means from two independent samples (groups) [two independent samples (groups) test, a form of the parallel-groups design in clinical trials].
3. Comparison of means from related samples (paired-sample tests).
4. One- and two-sample tests for proportions.
5. Tests to compare variances.

5.2 STATISTICAL HYPOTHESIS TESTING

To introduce the concept of hypothesis testing, we will use an example of the comparison of two treatment means (a two-sample test) that has many applications in pharmaceutical and clinical research. The details of the statistical test are presented in section 5.2.2. A clinical study is planned to compare the efficacy of a new antihypertensive agent to a placebo. Preliminary uncontrolled studies of the drug in humans suggest antihypertensive activity of the order of a drop of 10 to 15 mm Hg diastolic blood pressure. The proposed double- blind clinical trial is designed to study the effects of a once-a-day dose of tablets of the drug in a group of hypertensive patients. A second group of patients will receive an identical-appearing placebo.

Table 5.2 Average Results and Standard Deviation of a Clinical Study Comparing Drug and Placebo in the Treatment of Hypertension

	Drug	Placebo
Number of patients	11	10
Average blood pressure reduction (mm Hg)	10	1
Standard deviation	11.12	7.80

Blood pressure will be measured prior to the study and every two weeks after initiation of therapy for a total of eight weeks. For purposes of this presentation, we will be concerned only with the blood pressure at baseline (i.e., pretreatment) and after eight weeks of treatment. The variable that will be analyzed is the difference between the eight-week reading and the pretreatment reading. This difference, the change from baseline, will be called δ (delta). At the completion of the experiment, the average change from baseline will be compared for the active group and the placebo group in order to come to a decision concerning the efficacy of the drug in reducing blood pressure. The design is a typical *parallel-groups* design and the implementation of the study is straightforward. The problem, and question, that is of concern is: "What statistical techniques can be used to aid us in coming to a decision regarding the treatment (placebo and active drug) difference, and ultimately to a judgment of drug efficacy?"

From a qualitative and, indeed, practical point of view, a comparison of the *average* change in blood pressure for the active and placebo groups, integrated with previous experience, can give some idea of drug efficacy. Table 5.2 shows the average results of this study. (Only 21 patients completed the study.) Based on the results, our "internal computer" might reason as follows: "The new drug reduced the blood pressure by 10 mm Hg compared to a reduction of 1 mm Hg for patients on placebo. That is an impressive reduction for the drug"; or "The average reduction is quite impressive, but the sample size is small, less than 12 patients per group. If the raw data were available, it would be of interest to see how many patients showed an improvement when given the drug compared to the number who showed an improvement when given placebo." Particularly for small samples, one should examine the raw data. Such an examination of the clinical results may give an intuitive feeling of the effectiveness of a drug product. At one time, not very long ago, presentation of such experimental results accompanied by a subjective evaluation by the clinical investigator was important evidence in the support of efficacy of drugs. If the average results showed that the drug was no better than the placebo, the drug would probably be of little, if any interest.

One obvious problem with such a subjective analysis is the potential lack of consistency in the evaluation and conclusions that may be drawn from the same results by different reviewers. Also, although some experimental results may appear to point unequivocally to either efficacy or lack of efficacy, the inherent variability of the experimental data may be sufficiently large to obscure the truth. In general, subjective perusal of data is not sufficient to separate drug-related effects from random variability. In particular, comparing average results from small samples without a proper statistical analysis can be problematic. Statistical hypothesis testing is an objective means of assessing whether or not observed differences between treatments can be attributed to experimental variation (error). Good experimental design and data analysis are essential if clinical studies are to be used as evidence for drug safety and efficacy. This is particularly critical when such evidence is part of a New Drug Application (NDA) for the FDA, or for use for advertising claims.

The statistical evaluation or test of treatment differences is based on the *ratio* of the *observed treatment difference* (drug minus placebo in this example) to the *variability* of the difference. A large observed difference between drug and placebo accompanied by small variability is the most impressive evidence of a real drug effect (Fig. 5.6).

The magnitude of the ratio can be translated into a probability or "statistical" statement relating to the true but unknown drug effect. This is the basis of the common statement "statistically significant," implying that the difference observed between treatments is real, not merely a result of random variation. Statistical significance addresses the question of whether or not the treatments truly differ, but does not necessarily apply to the *practical* magnitude of the drug

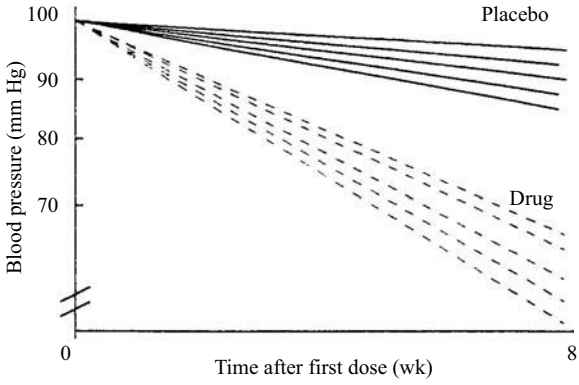


Figure 5.6 Mark of a real drug effect: a large difference between drug and placebo with small variation.

effect. The possibility exists that a small but real drug effect has no clinical meaning. Such judgments should be made by experts who can evaluate the magnitude of the drug effect in relation to the potential use of the drug vis-à-vis other therapeutic alternatives.

The preliminary discussion above suggests the procedure used in testing statistical hypotheses. Broadly speaking, data are first collected for comparative experiments according to an appropriate plan or design. For comparative experiments similar to that considered in our example, the *ratio* of the difference of the averages of the two treatments to its experimental error (standard deviation) is referred to an appropriate tabulated probability distribution. The treatment difference is deemed "statistically significant" if the ratio is sufficiently large relative to the tabulated probability values.

The testing procedure is based on the concept of a *null hypothesis.* The null hypothesis is a hypothetical statement about a parameter (such as the mean) that will subsequently be compared to the sample estimate of the parameter, to test for treatment differences. In the present example, the null hypothesis is

$$H_0 : \mu_1 = \mu_2 \quad \text{or} \quad \Delta = \mu_1 - \mu_2 = 0.$$

H_0 refers to the null hypothesis. μ_1 and μ_2 refer to the true blood pressure change from baseline for the two treatments. Δ is the hypothesized average difference of the change of blood pressure from baseline values for the new drug *compared* to placebo.

$\Delta =$ true average reduction in blood pressure due to drug minus true average reduction in blood pressure due to placebo

The sample estimate of Δ is designated as δ, and is assumed to have a normal distribution. The fact that H_0 is expressed as a specific difference (zero in this example), as opposed to a more general difference ($H_0 : \Delta \neq 0$), is an important concept. The test of "no difference" or some specific difference (e.g., $\Delta = 2$) is usually much more easily conceptualized and implemented than a test of some nonspecific difference.

The format of the null hypothesis statement is not always immediately apparent to those unfamiliar with statistical procedures. Table 5.3 shows some examples of how null hypothesis statements can be presented. The alternative hypothesis specifies alternative values of the parameter, which we accept as true if the statistical test leads to rejection of the null hypothesis. The alternative hypothesis includes values not specified in the null hypothesis. In our example, a reasonable alternative would include all values where the true values of the two means were not equal, typically stated as follows:

$$H_a : \mu_1 \neq \mu_2.$$

As noted above, the magnitude of the ratio of the (observed difference minus the hypothetical difference) to its variability, the s.d. of the observed difference, determines whether or

Table 5.3 Examples of the Null Hypothesis for Various Experimental Situations

Study	Null hypothesis	Comments
Effect of drug therapy on cholesterol level compared to placebo	$H_0 : \mu_1 = \mu_2$ or H_0: $\mu_1 - \mu_2 = 0$ or H_0:$\Delta = 0$	μ_1 refers to the true average cholesterol with drug and μ_1 refers to true average cholesterol with placebo
Effect of antibiotic on cure rate	$H_0 : p_0 = 0.8$	p_0 refers to the true proportion of patients cured; H_0 states that the hypothetical cure rate is 80%
Average tablet weight for quality control	$H_0 : w = 300$ mg	The target weight is a mean of 300 mg
Testing two mixing procedures with regard to homogeneity of the two mixes	$H_0 : \sigma_1^2 = \sigma_2^2$	The variance of the samples from the two procedures is hypothesized to be equal
Test to see if two treatments differ	$H_0 : \mu_1 \neq \mu_2$	This statement cannot be tested; H_0 must be specified as a specific difference or a limited range of differences

not H_0 should be accepted or rejected. A large ratio leads to rejection of H_0, and the difference is considered to be "statistically" significant. The specific details for testing simple hypotheses are presented below, beginning with the most elementary example, tests of a single mean.

5.2.1 Case I: Test of the Mean from a Single Population (One-Sample Tests), an Introduction to a Simple Example of Hypothesis Testing

The discussion above was concerned with a test to compare means from samples obtained from two groups, a drug group and a placebo group. The tests for a single mean are simpler in concept, and specific steps to construct this test are presented below. The process for other designs in which statistical hypotheses are tested is essentially the same as for the case described here. Other examples will be presented in the remainder of this chapter and, where applicable, in subsequent chapters of this book. The concept of hypothesis testing is important, and the student is well advised to make an extra effort to understand the procedures described below.

Data often come from a single population, and a comparison of the sample mean to some hypothetical or "standard" (known) value is desired. The examples shown in Table 5.4 are typical of those found in pharmaceutical research. The statistical test compares the observed value (a mean or a proportion, for example) to the hypothetical value.

To illustrate the procedure, we will consider an experiment to assess the effects of a change in manufacturing procedure on the average potency of a tablet product. A large amount of data was collected for the content of drug in the tablet formulation during a period of several years. The manufacturing process showed an average potency of 5.01 mg and a standard deviation of 0.11, both values considered to be equal to the true process parameters. A new batch was made with a modification of the usual manufacturing procedure. Twenty tablets were assayed

Table 5.4 Examples of Experiments Where a Single Population Mean Is Observed

Sample mean	Hypothetical or standard mean
Average tablet potency of *N* tablets	Label potency
Preference for product *A* in a paired preference test	50% are hypothesized to prefer product *A*
Average dissolution of *N* tablets	Quality control specifications
Proportion of patients cured by a new drug	Cure rate of *P*% based on previous therapy with a similar drug
Average cholesterol level of *N* patients under therapy	Hypothetical or standard value based on large amount of data collected by clinical laboratory
Average blood pressure reduction in *N* rats in preclinical study	Hypothetical average reduction considered to be of biological and clinical interest
Average difference of pain relief for two drugs taken by the same patients	Average difference (Δ) is hypothesized to be 0 if the drugs are identical

Table 5.5 Results of 20 Single-Tablet Assays from a Modification of a Process with a Historical Mean of 5.01 mg

5.13	5.04	5.09	5.00
4.98	5.03	5.01	4.99
5.20	5.08	4.96	5.18
5.08	5.06	5.02	5.24
4.99	5.17	5.06	5.00
	$\overline{X} = 5.0655$ mg	$S = 0.0806$	
	σ (historical) $= 0.11$		

and the results are shown in Table 5.5. The objective is to determine if the process modification results in a change of average potency from the process average of 5.01, the value of μ under the null hypothesis.

The steps for designing and analyzing this experiment are as follows:

1. *Careful planning* of the experiment ensures that the objectives of the experiment are addressed by an appropriate experimental design. The testing of a hypothesis where data are derived from a poorly implemented experiment can result in invalid conclusions. Proper design includes the choice and number of experimental units (patients, animals, tablets, etc.). Other considerations of experimental design and the manner in which observations are made are addressed in chapters 6, 8, and 11. Sample size may be determined on a scientific, statistical basis, but the choice is often limited by cost or time considerations, or the availability of experimental units. In the present example, the routine quality control content uniformity assay of 20 tablets was the determinant of sample size, a matter of convenience. The 20 tablets were chosen at random from the newly manufactured batch.
2. The *null hypothesis* and *alternative hypothesis* are defined prior to the implementation of the experiment or study. The usual test will be two sided

$$H_0 : \mu = \mu_0 \qquad H_a : \mu \neq \mu_0.$$

However, in the example below we will also discuss a one-sided test

$$H_0 : \mu \leq 5.01\,\text{mg} \qquad H_a : \mu > 5.01\,\text{mg}.$$

The objective of this experiment is to see if the average potency of the batch prepared with the modified procedure is different from that based on historical experience (5.01 mg). The null hypothesis takes the form of "no change," as discussed previously. To conclude that the new process has caused a change, we must demonstrate that the alternative hypothesis is true by rejecting the null hypothesis. The alternative hypothesis complements the null hypothesis. The two hypotheses are mutually exclusive and, together, in this example, cover all relevant possibilities that can result from the experiment. Either the average potency is 5.01 mg (H_0) or it is not (H_a). This is known as a *two-sided* (or *two-tailed*) test, suggesting that the average drug potency of the new batch can conceivably be smaller as well as greater than the historical process average of 5.01 mg. A *one-sided* test allows for the possibility of a difference in only one direction. Suppose that the process average of 5.01 mg suggested a preferential loss of drug during processing based on the theoretical amount added to the batch (e.g., 5.05 mg). The new procedure may have been designed to prevent this loss. Under these circumstances, one might hypothesize that the potency could only be greater (or, at least, not less) than the previous process average. Under this hypothesis, if the experiment reveals a lower potency than 5.01 mg, this result would be attributed to chance only; that is, although the average potency, in truth, is equal to or greater than 5.01 mg, chance variability may result in an experimental outcome where the observed average is "numerically" less than 5.01 mg. Such a result could occur, for example, as a result of a chance selection of

Table 5.6 Alpha and Beta Probabilities in Hypothesis Testing (Errors When Accepting or Rejecting H_0)

	H_0 is true	H_a (a specific alternative) is true
H_0 is rejected	Alpha (α)	1 − beta
H_0 is accepted	1 − alpha	Beta (β)

tablets of low potency for the assay sample. For a one-sided test, the null and alternative hypotheses may take the following form as noted above

$$H_0 : \mu \leq 5.01\,\text{mg} \qquad H_a : \mu > 5.01\,\text{mg}.$$

3. The *level of significance* is specified. This is the well-known *p* value associated with statements of statistical significance. The concept of the level of significance is crucial to an understanding of statistical methodology. The level of significance is defined as the probability that the statistical test results in a decision to reject H_0 (*a significant difference*) when, in fact, the treatments do not differ (H_0 is true). This concept will be clarified further when we describe the statistical test. By definition, the level of significance represents the chance of making a mistake when deciding to reject the null hypothesis. This mistake, or error, is also known as the *alpha* (α) *error* or error of the first kind (Table 5.6). Thus, if the statistical test results in rejection of the null hypothesis, we say that the difference is significant at the α level. If α is chosen to be 0.05, the difference is significant at the 5% level. This is often expressed, equivalently, as $p < 0.05$. Figure 5.7 shows values of $\overline{X}$ that lead to rejection of H_0 for a statistical test at the 5% level if σ is known.

   The *beta* (β) *error* is the probability of accepting H_0 (no treatment difference) when, in fact, some specified difference included in H_a is the true difference. Although the evaluation of the β error and its involvement in sample-size determination is important, because of the complex nature of this concept, further discussion of this topic will be delayed until chapter 6.

   The choice of magnitude of α, which should be established prior to the start of the experiment, rests on the experimenter or sponsoring organization. To make this choice, one should consider the risks or consequences that will result if an α error is made, that is, the error made when declaring that a significant difference exists when the treatments are indeed equivalent. Alpha should be defined prior to the experiment. It certainly would be unfair to choose an alpha after the results are obtained. Traditionally, α is chosen as 5% (0.05), although other levels such as 1% or 10% have been used. A justification for a level other than 5% should be forthcoming. An α error of 5% means that a decision that a significant difference exists (based on the rejection of H_0) has a probability of 5% (1 in 20) or less of being incorrect (*P less than or equal to* 0.05). Such a decision has credibility and is generally accepted as "proof" of a difference by regulatory agencies. When using the word "significant," one

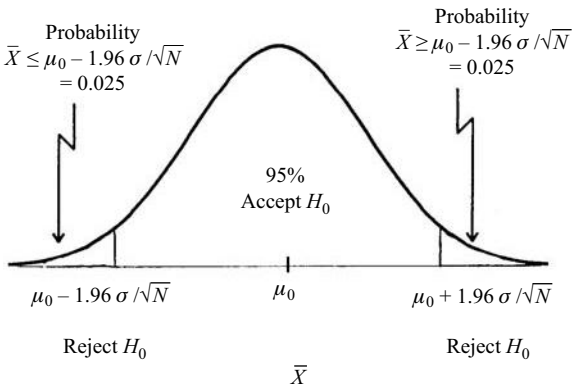


Figure 5.7 Region of rejection (critical region) in a statistical test (two-sided) at the 5% level with σ^2 known.

infers with a large degree of confidence that the experimental result does not support the null hypothesis.

An important concept is that if the statistical test results in a decision of *no significance*, the conclusion does *not* prove that H_0 is true or, in this case, that the average potency is 5.01 mg. Usually, "nonsignificance" is a weak statement, not carrying the clout or authority of the statement of "significance." Note that the chance of erroneously accepting H_0 is equal to β (Table 5.6). This means that β *percent* of the time, a nonsignificant result will be observed (H_0 is accepted as true), when a true difference specified by H_a or greater truly exists. Unfortunately, a good deal of the time when planning experiments, unlike α, β is not fixed in advance. The β level is often a result of circumstance. In most experiments, β is a consequence of the sample size, which is often based on considerations other than the size of β. However, the sample size is best computed with the aid of a predetermined value of β (see chap. 6). In our experiment, β was not fixed in advance. The sample of 20 tablets was chosen as a matter of tradition and convenience.

4. The *sample size*, in our example, has been fixed based on considerations that did not include β, as discussed above. However, the sample size can be calculated after α and β are specified, so that the experiment will be of sufficient size to have properties that will satisfy the choice of the α and β errors (see chap. 6 for further details).
5. After the experiment is completed, relevant statistics are computed. In this example and most situations with which we will be concerned, mean values are to be compared. It is at this point that the *statistical test of significance* is performed as follows. For a two-sided test, compute the ratio

$$Z = \frac{|\overline{X} - \mu_0|}{\sqrt{\sigma^2/N}} = \frac{|\overline{X} - \mu_0|}{\sigma/\sqrt{N}}. \tag{5.4}$$

The numerator of the ratio is the absolute value of the difference between the observed and hypothetical mean. (In a two-sided test, low or negative values as well as large positive values of the mean lead to significance.) The variance of $(\overline{X} - \mu_0)$[§] is equal to

$$\frac{\sigma^2}{N}.$$

The denominator of Eq. (5.4) is the standard deviation of the numerator. The Z ratio [Eq. (5.4)] consists of a difference, divided by its standard deviation. The ratio is exactly the Z transformation presented in chapter 3 [Eq. (3. 14)], which transforms a normal distribution with mean μ and variance σ^2 to the standard normal distribution ($\mu = 0$, $\sigma^2 = 1$).

In general, σ^2 is unknown, but it can be estimated from the sample data, and the sample estimate, S^2, is then used in the denominator of Eq. (5.4). An important question is how to determine if the ratio

$$t = \frac{|\overline{X} - \mu_0|}{\sqrt{S^2/N}} \tag{5.5}$$

leads to a decision of "significant." This prevalent situation (σ^2 unknown); will be discussed below.

As discussed above, significance is based on a probability statement defined by α. More specifically, the difference is considered to be statistically significant (H_0 is rejected) if the observed difference between the sample mean and μ_0 is sufficiently large so that the observed or larger differences are improbable (probability of α or less, e.g., $p \leq 0.05$) if the null hypothesis is true ($\mu = 5.01$ mg). In order to calculate the relevant probability, the observations are assumed to be statistically independent and normally distributed.

§ The variance of $(\overline{X} - \mu_0)$ is equal to the variance of $\overline{X}$ because μ_0 is constant and has a variance of 0.

With these assumptions, the ratio shown in Eq. (5.4) has a normal distribution with mean equal to 0 and variance equal to 1 (variance known, the standard normal distribution). The concept of the α error is illustrated in Figure 5.7. The values of $\overline{X}$ that lead to rejection of the null hypothesis define the "region of rejection," also known as the *critical region.* With a knowledge of the variance, the area corresponding to the critical region can be calculated using the standard normal distribution. The probability of observing a mean value in the critical region of the distribution defined by the null hypothesis is α. This region is usually taken as symmetrical areas in the tails of the distribution, with each tail containing $\alpha/2$ of the area ($2^1/_2$% in each tail at the 5% level) for a two-tailed test. Under the null hypothesis and the assumption of normality, $\overline{X}$ is normal with mean μ_0 and variance σ^2/N. The Z ratio [Eq. (5.4)] is a standard normal deviate, as noted above. Referring to Table IV.2, the values of $\overline{X}$ that satisfy

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{N}} \leq -1.96 \quad \text{or} \quad \frac{\overline{X} - \mu_0}{\sigma/\sqrt{N}} \geq +1.96 \tag{5.6}$$

will result in rejection of H_0 at the 5% level. The values of $\overline{X}$ that lead to rejection of H_0 may be derived by rearranging Eq. (5.6).

$$\overline{X} \leq \mu_0 - \frac{1.96\sigma}{\sqrt{N}} \quad \text{or} \quad \overline{X} \geq \mu_0 + \frac{1.96\sigma}{\sqrt{N}} \tag{5.7}$$

or, equivalently,

$$\left|\overline{X} - \mu_0\right| \geq \frac{1.96\sigma}{\sqrt{N}}. \tag{5.8}$$

If the value of $\overline{X}$ falls in the critical region, as defined in Eqs. (5.7) and (5.8), the null hypothesis is rejected and the difference is said to be significant at the α (5%) level.

The statistical test of the mean assay result from Table 5.5 may be performed: (a) assuming that σ is known ($\sigma = 0.11$) or (b) assuming that σ is unknown, but estimated from the sample ($S = 0.0806$).

The following examples demonstrate the procedure for applying the test of significance for a *single mean.*

(a) *One-sample test, variance known.* In this case, we believe that the large quantity of historical data defines the standard deviation of the process precisely, and that this standard deviation represents the variation in the new batch. We assume, therefore, that σ^2 is known. In addition, as noted above, if the data from the sample are independent and normally distributed, the test of significance is based on the standard normal curve (Table IV.2). The ratio as described in Eq. (5.4) is computed using the known value of the variance. If the absolute value of the ratio is greater than that which cuts off $\alpha/2$ percent of the area (defining the two tails of the rejection region, Fig. 5.7), the difference between the observed and hypothetical means is said to be significant at the α level. *For a two-sided test, the absolute value of the difference* is used because both large positive and negative differences are considered evidence for rejecting the null hypothesis.

In this example, we will use a two-sided test, because the change in potency, if any, may occur in either direction, higher or lower. The level of significance is set at the traditional 5% level.

$\alpha = 0.05$

Compute the ratio [Eq. (5.4)]

$$Z = \frac{\left|\overline{X} - \mu_0\right|}{\sigma/\sqrt{N}} = \frac{|5.0655 - 5.01|}{0.11/\sqrt{20}} = 2.26.$$

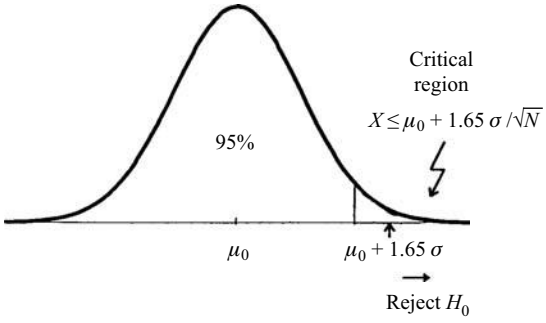


Figure 5.8 Rejection region for a one-sided test.

At the 5% level values of $|Z| \geq 1.96$ will lead to a declaration of significance for a two-sided test [Eq. (5.6)]. Therefore, the new batch can be said to have a potency different from previous batches (in this case, the mean is greater).

The level of significance is set *before* the actual experimental results are obtained. In the previous example, a one-sided test at the 5% level may be justified if convincing evidence were available to demonstrate that the new process would only result in mean results equal to or greater than the historical mean. If such a one-sided test had been deemed appropriate, the null hypothesis would be

$$H_0 : \mu = 5.01 \text{ mg}.$$

The alternative hypothesis, $H_a : \mu > 5.01$ mg, eliminates the possibility that the new process can lower the mean potency. The concept is illustrated in Figure 5.8. Now the rejection region lies only in values of $\overline{X}$ greater than 5.01 mg, as described below. An observed value of $\overline{X}$ below 5.01 mg is considered to be due only to chance (or it may be of no interest to us in other situations).

The rejection region is defined for values of $\overline{X}$ equal to or greater than $\mu_0 + 1.65\sigma/\sqrt{N}$ [or, equivalently, $(\overline{X} - \mu_0)/(\sigma/N) \geq 1.65$] because 5% of the area of the normal curve is found above this value (Table IV.2). This is in keeping with the definition of α: If the null hypothesis is true, we will erroneously reject the null hypothesis 5% of the time. Thus, we can see that a *smaller* difference is needed for significance using a one-sided test; the Z ratio need only exceed 1.65 rather than 1.96 for significance at the 5% level. In the present example, values of $\overline{X} \geq [5.01 + 1.65(0.11)/\sqrt{20}] = 5.051$ will lead to significance for a one-sided test. Clearly, the observed mean of 5.0655 is significantly different from 5.01 ($p < 0.05$). Note that in a one-sided test, the sign of the numerator is important and the absolute value is not used.

Usually, statistical tests are two-sided tests. One-sided tests are warranted in certain circumstances. However, the choice of a one-sided test should be made a priori, and one must be prepared to defend its use. As mentioned above, in the present example, if evidence were available to show that the new process could not reduce the potency, a one-sided test would be acceptable. To have such evidence and convince others (particularly, regulatory agencies) of its validity is not always an easy task. Also, from a scientific point of view, two-sided tests are desirable because significant results in both positive and negative directions are usually of interest.

(b) *One-sample test, variance unknown.* In most experiments in pharmaceutical research, the variance is unknown. Usually, the only estimate of the variance comes from the experimental data itself. As has been emphasized in the example above, use of the cumulative standard normal distribution (Table IV.2) to determine probabilities for the comparison of a mean to a known value (μ_0) is valid only if the variance is known.

The procedure for testing the significance of the difference of an observed mean from a hypothetical value (one-sample test) when the variance is estimated from the sample data is the same as that with the variance known, with the following exceptions:

1. The variance is computed from the experimental data. In the present example, the variance is $(0.0806)^2$; the standard deviation is 0.0806 from Table 5.5.

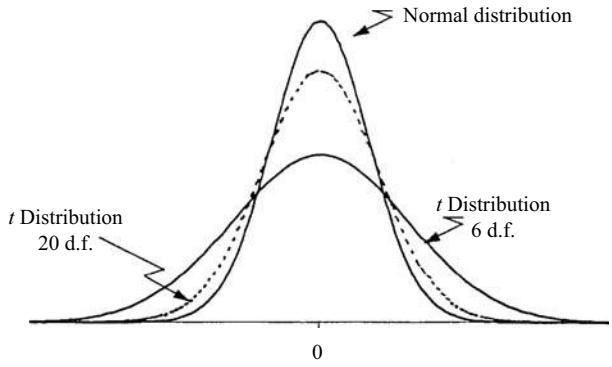


Figure 5.9 *t* distribution compared to the standard normal distribution.

2. The ratio is computed using S^2 instead of σ^2 as in Eq. (5.8a). This ratio

$$t = \frac{|\overline{X} - \mu_0|}{\sqrt{S^2/N}} \tag{5.8a}$$

is not distributed as a standard normal variable. If the mean is normally distributed, the ratio [Eq. (5.5)] has a *t* distribution. The *t* distribution looks like the standard normal distribution but has more area in the tails; the *t* distribution is more spread out. The shape of the *t* distribution depends on the d.f. As the d.f. increase the *t* distribution looks more and more like the standard normal distribution as shown in Figure 5.9. (Also, see sect. 3.5.2.) When the d.f. are equal to ∞ the *t* distribution is identical to the standard normal distribution (i.e., the variance is known).

The *t* distribution is a probability distribution that was introduced in section 5.1.1 and chapter 3. The area under the *t* distributions shown in Figure 5.9 is 1. Thus, as in the case of the normal distribution (or any continuous distribution), areas within specified intervals represent probabilities. However, unlike the normal distribution, there is *no* transformation that will change all *t* distributions (differing d.f.'s) to one "standard" *t* distribution. Clearly, a tabulation of all possible *t* distributions would be impossible. Table IV.4 shows commonly used probability points for representative *t* distributions. The values in the table are points in the *t* distribution representing cumulative areas (probabilities) of 80%, 90%, 95%, 97.5%, and 99.5%. For example, with d.f. = 10, 97.5% of the area of the *t* distribution is below a value of *t* equal to 2.23 (Fig. 5.10).

Note that when d.f. = ∞, the *t* value corresponding to a cumulative probability of 97.5% (0.975) is 1.96, exactly the same value as that for the standard normal distribution. Since the *t* distribution is symmetrical about zero, as is the standard normal distribution, a *t* value of -2.23 cuts off $1 - 0.975 = 0.025$ of the area (d.f. = 10). This means that to obtain a significant

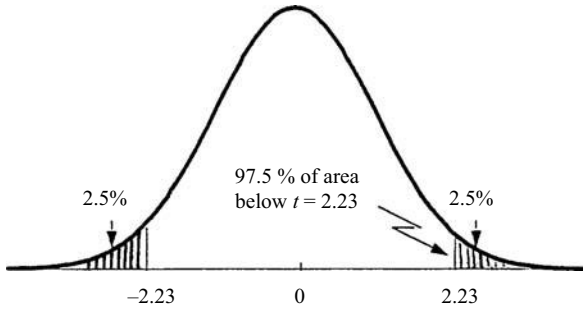


Figure 5.10 *t* distribution with 10 degrees of freedom.

difference of means at the 5% level for a two-sided test and d.f. equal to 10, the absolute value of the t ratio [Eq. (5.5)] must exceed 2.23. Thus the t values in the column headed "0.975" in Table IV.4 are values to be used for two-tailed significance tests at the 5% level (or for a two-sided 95% confidence interval). Similarly, the column headed "0.95" contains appropriate t values for significance tests at the 10% level for two-sided tests, or the 5% level for one-sided tests. The column headed "0.995" represents t values used for two-sided tests at the 1% level, or for 99% confidence intervals.

The number of d.f. used to obtain the appropriate value of t from Table IV.4 is the d.f. associated with the variance estimate in the denominator of the t ratio [Eq. (5.5)]. The d.f. for a mean are $N - 1$, or $19(20 - 1)$ in this example. The test is a two-sided test at the 5% level. The t ratio is

$$t = \frac{|\overline{X} - \mu_0|}{S/\sqrt{N}} = \frac{|5.0655 - 5.01|}{0.0806/\sqrt{20}} = 3.08.$$

The value of t needed for significance for a two-sided test at the 5% level is 2.09 (Table IV.4; 19 d.f.). Therefore, the new process results in a "significant" increase in potency ($p < 0.05$).

A 95% confidence interval for the true mean potency may be constructed as described in section 5.1.1 [Eq. (5.2)]

$$5.0655 \pm 2.09\left(\frac{0.0806}{\sqrt{20}}\right) = 5.028 \text{ to } 5.103 \text{ mg}.$$

Note that the notion of the *confidence interval* is closely associated with the *statistical test.* If the confidence interval covers the hypothetical value, the difference is not significant at the indicated level, and vice versa. In our example, the difference was significant at the 5% level, and the 95% confidence interval does *not* cover the hypothetical mean value of 5.01.

Example 4: As part of the process of new drug research, a pharmaceutical company places all new compounds through an "antihypertensive" screen. A new compound is given to a group of animals and the reduction in blood pressure measured. Experience has shown that a blood pressure reduction of more than 15 mm Hg in these hypertensive animals is an indication for further testing as a new drug candidate. Since such testing is expensive, the researchers wish to be reasonably sure that the compound truly reduces the blood pressure by more than 15 mm Hg before testing is continued; that is, they will continue testing only if the experimental evidence suggests that the true blood pressure reduction is greater than 15 mm Hg with a high probability.

$$H_o : \mu \leq 15\,\text{mm Hg reduction} \qquad H_a : \mu > 15\,\text{mm Hg reduction}$$

The null hypothesis is a statement that the new compound is unacceptable (blood pressure change is equal to or less than 15 mm Hg). This is typical of the concept of the null hypothesis. A rejection of the null hypothesis means that a difference probably exists. In our example, a true difference greater than 15 mm Hg means that the compound should be tested further. This is a *one-sided* test. Experimental results showing a difference of 15 mm Hg or less will result in a decision to accept H_0, and the compound will be put aside. If the blood pressure reduction exceeds 15 mm Hg the reduction will be tested for significance using a t test.

$$\alpha = 10\%(0.10)$$

The level of significance of 10% was chosen in lieu of the usual 5% level for the following reason. A 5% significance level means that 1 time in 20 a compound will be chosen as effective when the true reduction is less than 15 mm Hg. The company was willing to take a risk of 1 in 10 of following up an ineffective compound in order to reduce the risk of missing potentially effective compounds. One should understand that the choice of alpha and beta errors often is a compromise between reward and risk. We could increase the chances for reward, but we

Table 5.7 Blood Pressure Reduction Caused by a New Antihypertensive Compound in 10 Animals (mm Hg)

15	12
18	17
14	21
8	16
20	18
$\overline{X} = 15.9$	$S = 3.87$

could simultaneously increase the risk of failure, or, in this case, following up on an ineffective compound. Other things being equal, an increase in the α error decreases the β error; that is, there is a smaller chance of accepting H_0 when it is false. Note that the t value needed for significance is smaller at the 10% level than that at the 5% level. Therefore, a smaller reduction in blood pressure is needed for significance at the 10% level. The standard procedure in this company is to test the compound on 10 animals. The results shown in Table 5.7 were observed in a test of a newly synthesized potential antihypertensive agent.

The t test is [Eq. (5.5)]

$$t = \frac{15.9 - 15}{3.87/\sqrt{10}} = \frac{0.9}{1.22} = 0.74.$$

The value of t needed for significance is 1.38 (Table IV.4; one-sided test at the 10% level with 9 d.f.). Therefore, the compound is not sufficiently effective to be considered further. Although the average result was larger than 15 mm Hg, it was not sufficiently large to encourage further testing, according to the statistical criterion.

What difference (reduction) would have been needed to show a significant reduction, assuming that the sample variance does not change? Equation (5.5) may be rearranged as follows: $\overline{X} = t(S)/\sqrt{N} + \mu_0$. If $\overline{X}$ is greater than or equal to $t(S)/\sqrt{N} + \mu_0$, the average reduction will be significant, where t is the table value at the α level of significance with $(N - 1)$ d.f. In our example,

$$\frac{t(S)}{\sqrt{N}} + \mu_0 = \frac{(1.38)(3.87)}{\sqrt{10}} + 15 = 16.7.$$

A blood pressure reduction of 16.7 mm Hg or more (the critical region) would have resulted in a significant difference. (See Exercise Problem 10.)

5.2.2 Case II: Comparisons of Means from Two Independent Groups (Two Independent Groups Test)

A preliminary discussion of this test was presented in section 5.2. This most important test is commonly encountered in clinical studies (a parallel-groups design). Table 5.8 shows a few examples of research experiments that may be analyzed by the test described here. The data

Table 5.8 Some Examples of Experiments That May Be Analyzed by the Two-Independent- Groups Test

Clinical studies	Active drug compared to a standard drug or placebo; treatments given to different persons, one treatment per person
Preclinical studies	Comparison of drugs for efficacy and/or toxicity with treatments given to different animals
Comparison of product attributes from two batches	Tablet dissolution, potency, weight, etc., from two batches

of Table 5.2 will be used to illustrate this test. The experiment consisted of a comparison of an active drug and a placebo where each treatment is tested on different patients. The results of the study showed an average blood pressure reduction of 10 mm Hg for 11 patients receiving drug, and an average reduction of 1 mm Hg for 10 patients receiving placebo. The principal feature of this test (or design) is that treatments are given to two independent groups. The observations in one group are independent of those in the second group. In addition, we assume that the data within each group are normally and independently distributed.

The steps to be taken in performing the two independent groups test are similar to those described for the one-sample test (see sect. 5.2.1).

1. *Patients are randomly assigned to the two treatment groups.* (For a description of the method of random assignment, see chap. 4.) The number of patients chosen to participate in the study in this example was largely a consequence of cost and convenience. Without these restraints, a suitable sample size could be determined with a knowledge of β, as described in chapter 6. The drug and placebo were to be randomly assigned to each of 12 patients (12 patients for each treatment). There were several dropouts, resulting in 11 patients in the drug group and 10 patients in the placebo group.
2. *The null and alternative hypotheses are*

$$H_0\colon \mu_1 - \mu_2 = \Delta = 0 \qquad H_a : \Delta \neq 0.$$

   We hypothesize no difference between treatments. A "significant" result means that treatments are considered different. This is a two-sided test. The drug treatment may be better or worse than placebo.
3. α *is set at 0.05.*
4. *The form of the statistical test depends on whether or not variances are known.* In the usual circumstances, the variances are unknown.

5.2.2.1 Two Independent -Groups Test, Variances Known

If the variances of both groups are known, the ratio

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}} \tag{5.9}$$

has a normal distribution with mean 0 and standard deviation equal to 1 (the standard normal distribution). The numerator of the ratio is the difference between the observed difference of the means of the two groups ($\overline{X}_1 - \overline{X}_2$) and the hypothetical difference ($\mu_1 - \mu_2$ according to H_0). In the present case, and indeed in most of the examples of this test that we will consider, the hypothetical difference to be zero (i.e., $H_0 : \mu_1 - \mu_2 = 0$). The variability of $(\overline{X}_1 - \overline{X}_2)$¶ (defined as the standard deviation) is equal to

$$\sqrt{\sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2}}$$

[as described in App. I, if A and B are independent, $\sigma^2(A - B) = \sigma_A^2 + \sigma_B^2$]. Thus, as in the one-sample case, the test consists of forming a ratio whose distribution is defined by the standard normal curve. In the present example (test of an antihypertensive agent), suppose that the *variances* corresponding to drug and placebo are *known* to be 144 and 100, respectively. The rejection region is defined by α. For $\alpha = 0.05$, values of Z greater than 1.96 or less than -1.96 ($|Z| \geq 1.96$) will lead to rejection of the null hypothesis. Z is defined by Eq. (5.9).

¶ The variance of $(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)$ is equal to the variance of $(\overline{X}_1 - \overline{X}_2)$ because μ_1 and μ_2 are constants and have a variance equal to zero.

For a two-sided test

$$Z = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}},$$

$$\overline{X}_1 = 10, \quad \overline{X}_2 = 1, \quad N_1 = 11, \quad \text{and} \quad N_2 = 10.$$

Thus,

$$Z = \frac{|10 - 1|}{\sqrt{144/11 + 100/10}} = 1.87.$$

Since the absolute value of the ratio does not exceed 1.96, the difference is not significant at the 5% level. From Table IV.2, the probability of observing a value of Z greater than 1.87 is approximately 0.03. Therefore, the test can be considered significant at the 6% level [2(0.03) = 0.06 for a two-tailed test]. The probability of observing an absolute difference of 9 mm Hg or more between drug and placebo, if the two products are identical, is 0.06 or 6%.

We have set α equal to 5% as defining an unlikely event from a distribution with known mean (0) and variance (144/11 + 100/10 = 23.1). An event as far or farther from the mean (0) than 9 mm Hg can occur six times in a 100 if H_0 is true. Alternatively, the conclusion may be stated that the experimental results were not sufficient to reject H_0 because we set α at 5% a priori (i.e., before performing the experiment). In reality, there is nothing special about 5%. The use of 5% as the α level is based strongly on tradition and experience, as mentioned previously. Should significance at the 6% level result in a different decision than a level of 5%? To document efficacy, a significance level of 6% may not be adequate for acceptance by regulatory agencies. There has to be some cutoff point; otherwise, if 6% is acceptable, why not 7% and so on? However, for internal decisions or for leads in experiments used to obtain information for further work or to verify theories, 5% and 6% may be too close to "call." Rather than closing the door on experiments that show differences at $p = 0.06$, one might think of such results as being of "borderline" significance, worthy of a second look and/or further experimentation. In our example, had the difference between drug and placebo been approximately 9.4 mm Hg, we would have called the difference "significant," rejecting the hypothesis that the placebo treatment was equal to the drug.

P values are often presented with experimental results even though the statistical test shows nonsignificance at the predetermined α level. In this experiment, a statement that $p = 0.06$ ("The difference is significant at the 6% level") does not imply that the treatments are considered to be significantly different. We emphasize that if the α level is set at 5%, a decision that the treatments are different should be declared only if the experimental results show that $p \leq 0.05$. However, in practical situations, it is often useful for the experimenter and other interested parties to know the p value, particularly in the case of "borderline" significance.

5.2.2.2 Two- Independent -Groups Test, Variance Unknown

The procedure for comparing means of two independent groups when the variances are estimated from the sample data is the same as that with the variances known, with the following exceptions:

1. *The variance is computed from the sample data.* In order to perform the statistical test to be described below, in addition to the usual assumptions of normality and independence, *we assume that the variance is the same for each group.* (If the variances differ, a modified procedure can be used as described later in this chapter.) A rule of thumb for moderate-sized samples (N equal 10–20) is that the ratio of the two variances should not be greater than 3 to 4. Sometimes, in doubtful situations, a test for the equality of the two variances may be appropriate (see sect. 5.3) before performing the test of significance for means described here. To obtain an estimate of the common variance, first compute the variance of each group. The two

variances are *pooled* by calculating a weighted average of the variances, the best estimate of the true common variance. The weights are equal to the d.f., $N_1 - 1$ and $N_2 - 1$, for groups 1 and 2, respectively. N_1 and N_2 are the sample sizes for the two groups. The following formula may be used to calculate the pooled variance

$$S_p^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}. \tag{5.10}$$

Note that we do not calculate the pooled variance by first pooling together all of the data from the two groups. The pooled variance obtained by pooling the two *separate* variances will always be equal to or smaller than that computed from all of the data combined disregarding groups. In the latter case, the variance estimate includes the variability due to differences of means as well as that due to the variance within each group (see Exercise Problem 5). Appendix I has a further discussion of pooling variance.

2. *The ratio that is used for the statistical test is similar to Eq.* (5.9). Because the variance, S_p^2 (pooled variance), is estimated from the sample data, the ratio

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2/N_1 + S_p^2/N_2}} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/N_1 + 1/N_2}} \tag{5.11}$$

is used instead of Z [Eq. (5.9)]. The d.f. for the distribution are determined from the variance estimate, S_p^2. This is equal to the d.f., pooled from the two groups, equal to $(N_1 - 1) + (N_2 - 1)$ or $N_1 + N_2 - 2$.

These concepts are explained and clarified, step by step, in the following examples.

Example 5: Two different formulations of a tablet of a new drug are to be compared with regard to rate of dissolution. Ten tablets of each formulation are tested, and the percent dissolution after 15 minutes in the dissolution apparatus is observed. The results are tabulated in Table 5.9. The object of this experiment is to determine if the dissolution rates of the two formulations differ. The test for the "significance" of the observed difference is described in detail as follows:

1. *State the null and alternative hypotheses:*

$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$

Table 5.9 Percent Dissolution After 15 Minutes for Two Tablet Formulations

	Formulation *A*	**Formulation *B***
	68	74
	84	71
	81	79
	85	63
	75	80
	69	61
	80	69
	76	72
	79	80
	74	65
Average	77.1	71.4
Variance	33.43	48.71
s.d.	5.78	6.98

μ_1 and μ_2 are the true mean 15-minute dissolution values for formulations *A* and *B*, respectively. This is a two-sided test. There is no reason to believe that one or the other formulation will have a faster or slower dissolution, a priori.

2. *State the significance level* $\alpha = 0.05$. The level of significance is chosen as the traditional 5% level.
3. *Select the samples.* Ten tablets taken at random from each of the two pilot batches will be tested.
4. *Compute the value of the t statistic [Eq. (5.11)]:*

$$\frac{\left|\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)\right|}{S_p\sqrt{1/N_1 + 1/N_2}} = t = \frac{|77.1 - 71.4|}{S_p\sqrt{1/10 + 1/10}}$$

$\overline{X}_1 = 77.1$ and $\overline{X}_2 = 71.4$ (Table 5.9). $N_1 = N_2 = 10$ (d.f. = 9 for each group). S_p is calculated from Eq. (5.10)

$$S_p = \sqrt{\frac{9(33.43) + 9(48.71)}{18}} = 6.41.$$

Note that the *pooled standard deviation* is the *square root of the pooled variance*, where the pooled variance is a weighted average of the variances from each group. It is *not correct to average the standard deviations.* Although the sample variances of the two groups are not identical, they are "reasonably" close, close enough so that the assumption of equal variances can be considered to be acceptable. The assumption of equal variance and independence of the two groups is more critical than the assumption of normality of the data, because we are comparing means. Means tend to be normally distributed even when the individual data do not have a normal distribution, according to the central limit theorem. The observed value of *t* (18 d.f.) is

$$t = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{S_p\sqrt{1/N_1 + 1/N_2}} = \frac{|77.1 - 71.4|}{6.41\sqrt{2/10}} = 1.99.$$

Values of *t* equal to or greater than 2.10 (Table IV.4; d.f. = 18) lead to rejection of the null hypothesis. These values, which comprise the *critical region*, result in a declaration of "significance." In this experiment, the value of *t* is 1.99, and the difference is not significant at the 5% level ($p > 0.05$). This does not mean that the two formulations have the same rate of dissolution. The declaration of nonsignificance here probably means that the sample size was too small; that is, the same difference with a larger sample would be significant at the 5% level. Two different formulations are apt not to be identical with regard to dissolution. The question of statistical versus practical significance may be raised here. If the dissolutions are indeed different, will the difference of 5.7% (77.1–71.4%) affect drug absorption in vivo? A confidence interval on the difference of the means may be an appropriate way of presenting the results.

5.2.2.3 Confidence Interval for the Difference of Two Means

A confidence interval for the difference of two means can be constructed in a manner similar to that presented for a single mean as shown in section 5.1 [Eq. (5.2)]. For example, a confidence interval with a confidence coefficient of 95% is

$$(\overline{X}_1 - \overline{X}_2) \pm (t)S_p\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}, \tag{5.12}$$

t is the value obtained from Table IV.4 with appropriate d.f., with the probability used for a two-sided test. (Use the column labeled "0.975" in Table IV.4 for a 95% interval.) For the example

discussed above (tablet dissolution), a 95% confidence interval for the difference of the mean 15-minute dissolution values [Eq. (5.12)] is

$$(77.1 - 71.4) \pm 2.10(6.41)(0.447) = 5.7 \pm 6.02 = -0.32\% \text{ to } 11.72\%.$$

Thus the 95% confidence interval is from −0.32% to 11.72%.

5.2.2.4 Test of Significance If Variances of the Two Groups Are Unequal

If the two groups can be considered not to have equal variances and the variances are estimated from the samples, the usual t test procedure is not correct. This problem has been solved and is often denoted as the Behrens–Fisher procedure. Special tables are needed for the solution, but a good approximate test for the equality of two means can be performed using Eq. (5.13) [3].

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{S_1^2/N_1 + S_2^2/N_2}} \tag{5.13}$$

If $N_1 = N_2 = N$, then the critical t is taken from Table IV.4 with $N - 1$ instead of the usual $2(N - 1)$ d.f. If N_1 and N_2 are not equal, then the t value needed for significance is a weighted average of the appropriate t values from Table IV.4 with $N_1 - 1$ and $N_2 - 1$ d.f.

$$\text{Weighted average of } t \text{ values} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2},$$

where the weights are

$$w_1 = \frac{S_1^2}{N_1}, \qquad w_2 = \frac{S_2^2}{N_2}.$$

To make the calculation clear, assume that the means of two groups of patients treated with an antihypertensive agent showed the following reduction in blood pressure (mm Hg).

	Group *A*	**Group *B***
Mean	10.7	7.2
Variance (S^2)	51.8	5.3
N	20	15

We have reason to believe that the variances differ, and for a two-sided test, we first calculate t' according to Eq. (5.13)

$$t' = \frac{|10.7 - 7.2|}{\sqrt{51.8/20 + 5.3/15}} = 2.04.$$

The critical value of t' is obtained using the weighting procedure. At the 5% level, t with 19 d.f. $= 2.09$ and t with 14 d.f. $= 2.14$. The weighted average t value is

$$\frac{(51.8/20)(2.09) + (5.3/15)(2.14)}{(51.8/20) + (5.3/15)} = 2.10.$$

Since t' is less than 2.10, the difference is considered to be not significant at the 5% level.

5.2.2.5 Overlapping Confidence Intervals and Statistical Significance

When comparing two independent treatments for statistical significance, sometimes people erroneously make conclusions based on the confidence intervals constructed from each treatment separately. In particular, if the confidence intervals overlap, the treatments are considered not to differ. This reasoning is not necessarily correct. The fallacy can be easily seen from the following example. Consider two independent treatments, A and B, representing two formulations of the same drug with the following dissolution results:

Treatment	N	Average	s.d.
A	6	37.5	6.2
B	6	47.4	7.4

For a two-sided test, the two-sample t test results in a t value of

$$t = \frac{|47.4 - 37.5|}{6.83\sqrt{1/6 + 1/6}} = 2.51.$$

Since 2.51 exceeds the critical t value with l0 d.f. (2.23), the results show significance at the 5% level.

Computation of the 95% confidence intervals for the two treatments results in the following:

Treatment A : $37.5 \pm (2.57)(6.2)\sqrt{1/6} = 30.99$ to 44.01.
Treatment B : $47.4 \pm (2.57)(7.4)\sqrt{1/6} = 39.64$ to 55.16.

Clearly, in this example, the individual confidence intervals overlap (the values between 39.64 and 44.01 are common to both intervals), yet the treatments are significantly different. The 95% confidence interval for the difference of the two treatments is

$$(47.4 - 37.5) \pm 8.79 = 1.1 \text{ to } 18.19.$$

As has been noted earlier in this section, if the 95% confidence interval does not cover 0, the difference between the treatments is significant at the 5% level.

5.2.2.6 Summary of t-Test Procedure and Design for Comparison of Two Independent Groups

The t-test procedure is essentially the same as the test using the normal distribution (Z test). The t test is used when the variance(s) are unknown and estimated from the sample data. The t distribution with ∞ d.f. is identical to the standard normal distribution. Therefore, the t distribution with ∞ d.f. can be used for normal distribution tests (e.g., comparison of means with variance known). When using the t test, it is necessary to compute a pooled variance. [With variances known, a pooled variance is not computed; see Eqs. (5.10) and (5.11).] An assumption underlying the use of this t test is that the variances of the comparative groups are the same. Other assumptions when using the t test are that the data from the two groups are independent and normally distributed. If the variances are considered to be unequal, use the approximate Behrens–Fisher method.

If H_0 is rejected (the difference is "significant"), one accepts the alternative, $H_a := \mu_1 \neq \mu_2$ or $\mu_1 - \mu_2 \neq 0$. The best estimate of the true difference between the means is the observed difference. A confidence interval gives a range for the true difference (see above). If the confidence interval covers 0, the statistical test is not significant at the corresponding alpha level.

Planning an experiment to compare the means of two independent groups usually requires the following considerations:

1. *Define the objective.* For example, in the example above, the objective was to determine if the two formulations differed with regard to rates of dissolution.

2. Determine the *number of samples* (experimental units) to be included in the experiment. We have noted that statistical methods may be used to determine the sample size (chap. 6). However, practical considerations such as cost and time constraints are often predominating factors. The *sample size* of the two groups *need not be equal* in this type of design, also known as a *parallel-groups* or *one-way analysis of variance* design. If the primary interest is the comparison of means of the two groups, equal sample sizes are optimal (assuming that the variances of the two groups are equal). That is, given the total number of experimental units available (patients, tablets, etc.), the most powerful comparison will be obtained by dividing the total number of experimental units into two equal groups. The reason for this is that $(1/N_1) + (1/N_2)$, which is in the denominator of the test ratio, is minimal when $N_1 = N_2 = N_t/2$ (N_t is the total sample size). In many circumstances (particularly in clinical studies), observations are lost due to errors, patient dropouts, and so on. The analysis described here is still valid, but some power will be lost. *Power* is the ability of the test to discriminate between the treatment groups. (Power is discussed in detail in chap. 6.) Sometimes, it is appropriate to use different sample sizes for the two groups. In a clinical study where a new drug treatment is to be compared to a standard or placebo treatment, one may wish to obtain data on adverse experiences due to the new drug entity in addition to comparisons of efficacy based on some relevant mean outcome. In this case, the design may include more patients on the new drug than the comparative treatment. Also, if the variances of two groups are known to be unequal, the optimal sample sizes will not be equal [4].
3. *Choose the samples.* It would seem best in many situations to be able to apply treatments to randomly chosen experimental units (e.g., patients). Often, practical considerations make this procedure impossible, and some compromise must be made. In clinical trials, it is usually not possible to select patients at random according to the strict definition of "random." We usually choose investigators who assign treatments to the patients available to the study in a random manner.
4. *Observations are made* on the samples. Every effort should be made to avoid bias. Blinding techniques and randomizing the order of observations (e.g., assays) are examples of ways to avoid bias. Given a choice, objective measurements, such as body weights, blood pressure, and blood assays, are usually preferable to subjective measurements, such as degree of improvement, psychological traits, and so on.
5. The *statistical analysis*, as described above, is then applied to the data. The statistical methods and probability levels (e.g., α) should be established prior to the experiment. However, one should not be immobilized because of prior commitments. If experimental conditions differ from that anticipated, and alternative analyses are warranted, a certain degree of flexibility is desirable. However, statistical theory (and common sense) shows that it is not fair to examine the data to look for all possible effects not included in the objectives. The more one looks, the more one will find. In a large data set, any number of unusual findings will be apparent if the data are examined with a "fine-tooth comb," sometimes called "data dredging." If such unexpected results are of interest, it is best to design a new experiment to explore and define these effects. Otherwise, large data sets can be incorrectly used to demonstrate a large number of unusual, but inadvertent, random, and inconsequential "statistically" significant differences.

5.2.3 Test for Comparison of Means of Related Samples (Paired-Sample *t* Test)

Experiments are often designed so that comparisons of two means are made on related samples. This design is usually more *sensitive* than the two independent groups t test. A test is more sensitive if the experimental variability is smaller. With smaller variability, smaller differences can be detected as statistically significant. In clinical studies, a paired design is often described as one in which each patient acts as his or her own "control." A bioequivalence study, in which each subject takes each of a test and reference drug product, is a form of paired design (see sect. 11.4).

In the paired-sample experiment, the two treatments are applied to experimental units that are closely related. If the same person takes both treatments, the relationship is obvious. Table 5.10 shows common examples of related samples used in paired tests.

Table 5.10 Examples of Related Samples

Clinical studies	Each patient takes each drug on different occasions (e.g., crossover study)
	Each patient takes each drug simultaneously, such as in skin testing; for example, an ointment is applied to different parts of the body
	Matched pairs: two patients are matched for relevant characteristics (age, sex, disease state, etc.) and two drugs randomly assigned, one to each patient
Preclinical studies	Drugs assigned randomly to littermates
Analytical development	Same analyst assays all samples
	Each laboratory assays all samples in collaborative test
	Each method is applied to a homogeneous sample
Stability studies	Assays over time from material from same container

The paired t test is identical in its implementation to the one-sample test described in section 5.2.1. In the paired test, the single sample is obtained by taking differences between the data for the two treatments for each experimental unit (patient or subject, for example). With N pairs of individuals, there are N data points (i.e., N differences). The N differences are designated as δ. Example 4, concerning the average reduction in blood pressure in a preclinical screen, was a paired-sample test in disguise. The paired data consisted of pre- and postdrug blood pressure readings for each animal. We were interested in the difference of pre- and postvalues (δ), the blood pressure reduction (see illustration below).

In paired tests, treatments should be assigned either in random order, or in some designed way, as in the crossover design. In the crossover design, usually one-half of the subjects receive the two treatments in the order *A*-*B*, and the remaining half of the subjects receive the treatments in the opposite order, where *A* and *B* are the two treatments. The crossover design is discussed in detail in chapter 11. With regard to blood pressure reduction, it is obvious that the order cannot be randomized. The pretreatment reading occurs before the post-treatment reading. The inflexibility of this ordering can create problems in interpretation of such data. The conclusions based on these data could be controversial because of the lack of a "control" group. If extraneous conditions that could influence the experimental outcome are different at the time of the initial and final observation (pre- and post-treatment), the treatment effect is "confounded" with the differences in conditions at the two points of observation. Therefore, randomization of the order of treatment given to each subject is important for the validity of this statistical test. For example, consider a study to compare two hypnotic drugs with regard to sleep-inducing effects. If the first drug were given to all patients before the second drug, and the initial period happened to be associated with hot and humid weather conditions, any observed differences between drugs (or lack of difference) would be "tainted" by the effect of the weather on the therapeutic response.

An important feature of the paired design is that the experimental units receiving the two treatments are, indeed, related. Sometimes, this is not as obvious as the example of the same

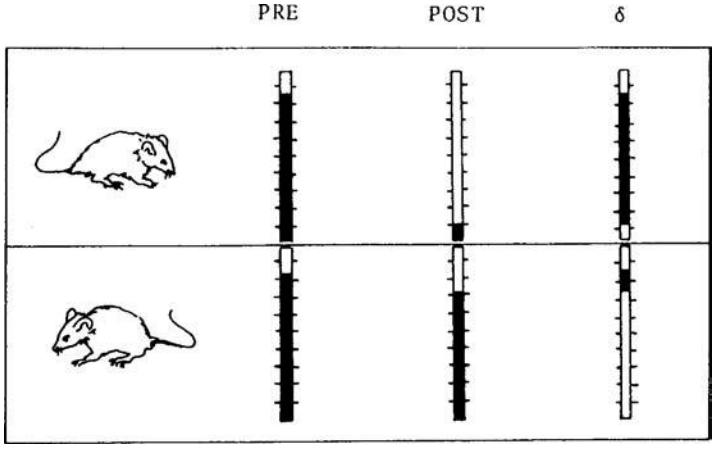


Table 5.11 Results of a Bioavailability Study Comparing a New Formulation (*A*) to a Marketed Form (*B*) with Regard to the Area Under the Blood-Level Curve

Animal	*A*	*B*	$\delta = B - A$	$A/B = R$
1	136	166	30	0.82
2	168	184	16	0.91
3	160	193	33	0.83
4	94	105	11	0.90
5	200	198	−2	1.01
6	174	197	23	0.88
			$\bar{\delta} = 18.5$	$\bar{R} = 0.89$
			$S_\delta = 13.0$	$S_R = 0.069$

patient taking both treatments. One can think of the concept of relatedness in terms of the paired samples being more alike than samples from members of different pairs. Pairs may be devised in clinical trials by pairing patients with similar characteristics, such as age, sex, severity of disease, and so on.

Example 6: A new formulation of a marketed drug is to be tested for bioavailability, comparing the extent of absorption to the marketed form on six laboratory animals. Each animal received both formulations in random order on two different occasions. The results, the area under the blood level versus time curve (AUC), are shown in Table 5.11.

$$H_0 : \Delta = 0^{**} \qquad H_a : \ \Delta \neq 0$$

This is a two-sided test, with the null hypothesis of equality of means of the paired samples. (The true difference is zero.) Before the experiment, it was not known which formulation would be more or less bioavailable if, indeed, the formulations are different. The significance level is set at 5%. From Table 5.11, the average difference is 18.5 and the standard deviation of the differences (δ values) is 13.0. The *t* test is

$$t = \frac{\bar{\delta} - \Delta}{S/\sqrt{N}}. \tag{5.14}$$

The form of the test is the same as the one-sample *t* test [Eq. (5.5)]. In our example, a two-sided test,

$$t = \frac{|18.5 - 0|}{13/\sqrt{6}} = 3.84.$$

For a two-sided test at the 5% level, a *t* value of 2.57 is needed for significance (d.f. = 5; there are six pairs). Therefore, the difference is significant at the 5% level. Formulation *B* appears to be more bioavailable.

In many kinds of experiments, *ratios* are more meaningful than differences as a practical expression of the results. In comparative bioavailability studies, the ratio of the AUCs of the two competing formulations is more easily interpreted than is their difference. The ratio expresses the *relative* absorption of the formulations. From a statistical point of view, if the AUCs for formulations *A* and *B* are normally distributed, the difference of the AUCs is also normally distributed. It can be proven that the ratio of the AUCs will *not* be normally distributed and the assumption of normality for the *t* test is violated. However, if the variability of the ratios is not great and the sample size is sufficiently "large," analyzing the ratios should give conclusions similar to that obtained from the analysis of the differences. Another alternative for the analysis of such data is the logarithmic transformation (see chap. 10), where the differences of the

** Δ is the hypothetical difference and δ the observed average difference.

logarithms of the AUCs are analyzed. (Also see chap. 11, Locke's approach, for an analysis of ratios.) For purposes of illustration, we will analyze the data in Table 5.11 using the ratio of the AUCs for formulations A and B. The ratios are calculated in the last column in Table 5.11.

The null and alternative hypotheses in this case are

$$H_0 : R_0 = 1 \qquad H_a : R_0 \neq 1,$$

where R_0 is the true ratio. If the products are identical, we would expect to observe an average ratio close to 1 from the experimental data. For the statistical test, we chooser α equal to 0.05 for a two-sided test. Applying Eq. (5.5), where $\overline{X}$ is replaced by the average ratio $\overline{R}$

$$t = \frac{|\overline{R} - 1|}{S/\sqrt{6}} = \frac{|0.89 - 1|}{0.069/\sqrt{6}} = 3.85.$$

Note that this is a one-sample test. We are testing the mean of a single sample of ratios versus the hypothetical value of 1. Because this is a two-sided test, low or high ratios can lead to significant differences. As in the analysis of the differences, the value of t is significant at the 5% level. (According to Table IV.4, at the 5% level, t must exceed 2.57 for significance.)

A *confidence interval* for the average ratio (or difference) of the AUCs can be computed in a manner similar to that presented earlier in this chapter [Eq. (5.2)]. A 95% confidence interval for the true ratio A/B is given by

$$\frac{A}{B} = \overline{R} \pm \frac{t(S)}{\sqrt{N}} = 0.89 \pm \frac{2.57(0.069)}{\sqrt{6}} = 0.89 \pm 0.07 = 0.82 \text{ to } 0.96.$$

Again, the fact that the confidence interval does not cover the value specified by H_0 (1) means that the statistical test is significant at the 5% level.

A more complete discussion of the analysis of bioequivalence data as required by the FDA is given in chapter 11.

5.2.4 Normal Distribution Tests for Proportions (Binomial Tests)

The tests described thus far in this chapter (normal distribution and t tests as well as confidence intervals) can also be applied to data that are binomially distributed. To apply tests for binomial variables based on the normal distribution, a conservative rule is that the sample sizes should be sufficiently large so that both $N\hat{p}$ and $N\hat{q}$ are larger than or equal to 5. Where $\hat{p}$ is the observed proportion and $\hat{q} = 1 - \hat{p}$. For symmetric distributions ($p \cong 0.5$), this constraint may be relaxed somewhat. The binomial tests are based on the normal approximation to the binomial and, therefore, we use normal curve probabilities when making decisions in these tests. To obtain the probabilities for tests of significance, we can use the t table with ∞ d.f. or the standard normal distribution (Tables IV.4 and IV.2, respectively). We will also discuss the application of the χ^2 (chi-square) distribution to the problem of comparing the "means" of binomial populations.

5.2.4.1 Test to Compare the Proportion of a Sample to a Known or Hypothetical Proportion

This test is equivalent to the normal test of the mean of a single population. The test is

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / N}} \tag{5.15}$$

where $\hat{p}$ is the observed proportion and p_0 the hypothetical proportion under the null hypothesis H_0:$p' = p_0$.

The test procedure is analogous to the one-sample tests described in section 5.2.1. Because of the discrete nature of binomial data, a correction factor is recommended to improve the

normal approximation. The correction, often called the *Yates continuity correction*, consists of subtracting $1/(2N)$ from the absolute value of the numerator of the test statistic [Eq. (5.15)]

$$Z = \frac{|\hat{p} - p_0| - 1/(2N)}{\sqrt{p_0 q_0 / N}}. \tag{5.16}$$

For a two-tailed test, the approximation can be improved as described by Snedecor and Cochran [5]. The correction is the same as the Yates correction if np is "a whole number or ends in 0.5." Otherwise, the correction is somewhat less than $1/(2N)$ (see Ref. [5] for details). In the examples presented here, we will use the Yates correction. This results in probabilities very close to those that would be obtained by using exact calculations based on the binomial theorem. Some examples should make the procedure clear.

Example 7: Two products are to be compared for preference with regard to some attribute. The attribute could be sensory (taste, smell, etc.) or therapeutic effect as examples. Suppose that an ointment is formulated for rectal itch and is to be compared to a marketed formulation. Twenty patients try each product under "blind" conditions and report their preference. The null hypothesis and alternative hypothesis are

$$H_0: p_a = p_b \quad \text{or} \quad H_0 : p_a = 0.5 \qquad \dot{H}_a : p_a \neq 0.5,$$

where p_a and p_b are the hypothetical preferences for A and B, respectively. If the products are truly equivalent, we would expect one-half of the patients to prefer either product A or B. Note that is a *one-sample* test. There are two possible outcomes that can result from each observation: a patient prefers A or prefers B ($p_a + p_b = 1$).

We observe the proportion of preferences (successes) for A, where A is the new formulation. This is a two-sided test; very few or very many preferences for A would suggest a significant difference in preference for the two products. Final tabulation of results showed that 15 of 20 patients found product A superior (5 found B superior). Does this result represent a "significant" preference for product A? Applying Eq. (5.16), we have

$$Z = \frac{|15/20 - 0.5| - 1/40}{\sqrt{(0.5)(0.5)/20}} = 2.01.$$

Note the correction for continuity, $1/(2N)$. Also note that the denominator uses the value of pq based on the null hypothesis ($p_a = 0.5$), not the sample proportion ($0.75 = 15/20$). This procedure may be rationalized if one verbalizes the nature of the test. We assume that the preferences are equal for both products ($p_a = 0.5$). We then observe a sample of 20 patients to see if the results conform with the hypothetical preference. Thus, the test is based on a hypothetical binomial distribution with the expected number of preferences equal to 10 ($p_a \times 20$). See Figure 5.11, which illustrates the rejection region in this test. The value of $Z = 2.01$ (15 preferences in a sample of 20) is sufficiently large to reject the null hypothesis. A value of 1.96 or greater is

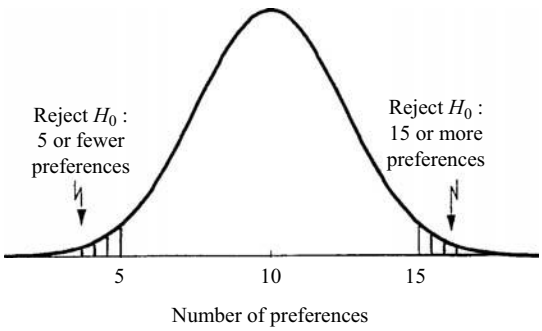


Figure 5.11 Rejection region for the test of $p_a = 0.5$ for a sample of 20 patients ($\alpha = 0.05$, two-sided test).

significant at the 5% level (Table IV.2). The test of $p_0 = 0.5$ is common in statistical procedures. The sign test described in chapter 15 is a test of equal proportions (i.e., $p_0 = q_0 = 0.5$).

Example 8: A particularly lethal disease is known to result in 95% fatality if not treated. A new treatment is given to 100 patients and 10 survive. Does the treatment merit serious consideration as a new therapeutic regimen for the disease? We can use the normal approximation because the expected number of successes and failures both are ≥ 5, that is, $Np_0 = 5$ and $Nq_0 = 5$ and $Nq_0 = 95$ ($p_0 = 0.05$, $N = 100$). A one-sided test is performed because evidence supports the hypothesis that the treatment cannot worsen the chances of survival. The α level is set at 0.05. Applying Eq. (5.16), we have

$$H_0 : p_0 = 0.05 \qquad H_0 : p_0 > 0.05$$

$$Z = \frac{|0.10 - 0.05| - 1/200}{\sqrt{(0.05)(0.95)/100}} = 2.06.$$

Table IV.2 shows that a value of Z equal to 1.65 would result in significance at the 5% level (one-sided test). Therefore, the result of the experiment is strong evidence that the new treatment is effective ($p < 0.05$).

If either Np_0 or Nq_0 is less than 5, the normal approximation to the binomial may not be justified. Although this rule is conservative, if in doubt, in these cases, probabilities must be calculated by enumerating all possible results that are equally or less likely to occur than the observed result under the null hypothesis. This is a tedious procedure, but in some cases it is the only way to obtain the probability for significance testing [7]. Fortunately, most of the time, the sample sizes of binomial experiments are sufficiently large to use the normal approximation.

5.2.4.2 Tests for the Comparison of Proportions from Two Independent Groups

Experiments commonly occur in the pharmaceutical and biological sciences that involve the comparison of proportions from two independent groups. These experiments are analogous to the comparison of means in two independent groups using the t or normal distributions. For proportions, the form of the test is similar. With a sufficiently large sample size, the normal approximation to the binomial can be used, as in the single-sample test. For the hypothesis: $H_0 : p_a = p_b (p_a - p_b = 0)$, the test using the normal approximation is

$$Z = \frac{\hat{p}_a - \hat{p}_b}{\sqrt{\hat{p}_0 \hat{q}_0 (1/N_1 + 1/N_2)}}, \tag{5.17}$$

where $\hat{p}_a$ and $\hat{p}_b$, are the observed proportions in groups A and B, respectively, and N_1 and N_2 are the sample sizes for groups A and B, respectively, $\hat{p}_0$ and $\hat{q}_0$ are the "pooled" proportion of successes and failures. The pooled proportion, $\hat{p}_0$, is similar to the pooled standard deviation in the t test. For proportions, the results of the two comparative groups are pooled together and the "overall" observed proportion is equal to $\hat{p}_0$. Under the null hypothesis, the probability of success is the same for both groups, A and B. Therefore, the best estimate of the common probability for the two groups is the estimate based on the combination of data from the entire experiment. An example of this calculation is shown in Table 5.12. The pooled proportion, $\hat{p}_0$, is a weighted average of the two proportions. This is exactly the same as adding up the total number of "successes" and dividing this by the total number of observations. In the example in

Table 5.12 Sample Calculation for Pooling Proportions from Two Groups

Group I	Group II
$N = 20$	$N = 30$
$\hat{p}_1 = 0.8$	$\hat{p}_2 = 0.6$
$\hat{p}_0 =$ pooled $p = (20 \times 0.8 + 30 \times 0.6)/(20 + 30) = 0.68$	

Table 5.12, the total number of successes is 34, 16 in group I and 18 in group II. The total number of observations is 50, 30 + 20. The following examples illustrate the computations.

Example 9: In a clinical study designed to test the safety and efficacy of a new therapeutic agent, the incidence of side effects are compared for two groups of patients, one taking the new drug and the other group taking a marketed standard agent. Headache is a known side effect of such therapy. Of 212 patients on the new drug, 35 related that they had experienced severe headaches. Of 196 patients on the standard therapy, 46 suffered from severe headaches. Can the new drug be claimed to result in fewer headaches than the standard drug at the 5% level of significance? The null and alternative hypotheses are

$$H_0 : p_1 = p_2 (p_1 - p_2) = 0 \qquad H_a : p_1 \neq p_2.$$

This is a two-sided test. Before performing the statistical test, the following computations are necessary

$$\hat{p}_1 = \frac{35}{212} = 0.165$$
$$\hat{p}_2 = \frac{46}{196} = 0.235$$
$$\hat{p}_0 = \frac{81}{408} = 0.199 \, (\hat{q}_0 = 0.801).$$

Applying Eq. (5.17), we have

$$Z = \frac{|0.235 - 0.165|}{\sqrt{(0.199)(0.801)(1/212 + 1/196)}} = \frac{0.07}{0.0395} = 1.77.$$

Since a Z value of 1.96 is needed for significance at the 5% level, the observed difference between the two groups with regard to the side effect of "headache" is not significant ($p > 0.05$).

Example 10: In a preclinical test, the carcinogenicity potential of a new compound is determined by administering several doses to different groups of animals. A control group (placebo) is included in the study as a reference. One of the dosage groups showed an incidence of the carcinoma in 9 of 60 animals (15%). The control group exhibited 6 carcinomas in 65 animals (9.2%). Is there a difference in the proportion of animals with the carcinoma in the two groups ($\alpha = 5\%$)? Applying Eq. (5.17), we have

$$H_0 : p_1 = p_2 \qquad H_0 : p_1 \neq p_2$$

$$Z = \frac{|9/60 - 6/65|}{\sqrt{(15/125)(110/125)(1/60 + 1/65)}} = \frac{0.0577}{0.058} = 0.99.$$

Note that $\hat{p}_1 = 9/60 = 0.15$, $\hat{p}_2 = 6/65 = 0.092$, and $\hat{p}_0 = 15/125 = 0.12$.

Since Z does not exceed 1.96, the difference is not significant at the 5% level. This test could have been a one-sided test (a priori) if one were certain that the new compound could not lower the risk of carcinoma. However, the result is not significant at the 5% level for a one-sided test; a value of Z equal to 1.65 or greater is needed for significance for a one-sided test.

Example 11: A new operator is assigned to a tablet machine. A sample of 1000 tablets from this machine showed 8% defects. A random sample of 1000 tablets from the other tablet presses used during this run showed 5.7% defects. Is there reason to believe that the new operator produced more defective tablets than that produced by the more experienced personnel? We will perform a two-sided test at the 5% level, using Eq. (5.17).

$$Z = \frac{|0.08 - 0.057|}{\sqrt{(0.0685)(0.9315)(2/1000)}} = \frac{0.023}{0.0113} = 2.04$$

Since the value of Z (2.04) is greater than 1.96, the difference is significant at the 5% level. We can conclude that the new operator is responsible for the larger number of defective tablets produced at his station, assuming that there is no difference among tablet presses. (See also Exercise Problem 19.) If a continuity correction is used, the equivalent chi-square test with a correction as described below is recommended.**

There is some controversy about the appropriateness of a continuity correction in these tests. D'Agostino et al. [6] examined various alternatives and compared the results to exact probabilities. They concluded that for small sample sizes ($N_1$and $N_2 < 15$), the use of the Yates continuity correction resulted in too conservative probabilities (i.e., probabilities were too high which may lead to a lack of rejection of H_0 in some cases).

They suggest that in these situations a correction should not be used. They also suggest an alternative analysis that is similar to the t test

$$t = \frac{|p_1 - p_2|}{\text{s.d.}\sqrt{1/N_1 + 1/N_2}}, \tag{5.18}$$

where s.d. is the pooled standard deviation computed from the data considering a success equal to 1 and a failure equal to 0. The value of t is compared to the appropriate t value with $N_1 + N_2 - 2$ d.f. The computation for the example in Table 5.12 is as follows:

For group I, $S_1^2 = \frac{16-(16^2/20)}{19} = 0.168$.
For group II, $S_2^2 = \frac{18-(18^2/30)}{29} = 0.248$.

(Note that for group I the number of successes is 16 and the number of failures is 4. Thus, we have 16 values equal to 1 and 4 values equal to 0. The variance is calculated from these 20 values.) The pooled variance is

$$\frac{19 \times 0.168 + 29 \times 0.248}{48} = 0.216.$$

The pooled standard deviation is 0.465.
From Eq. (5.18),

$$t = \frac{|0.8 - 0.6|}{0.465\sqrt{1/20 + 1/30}} = 1.49.$$

The t value with 48 d.f. for significance at the 5% level for a two-sided test is 2.01. Therefore, the results fail to show a significant difference at the 5% level.

Fleiss [7] advocates the use of the Yates continuity correction. He states "Because the correction for continuity brings probabilities associated with χ^2 and Z into close agreement with the exact probabilities, the correction should always be used."

5.2.5 Chi-Square Tests for Proportions

An alternative method of comparing proportions is the chi-square (χ^2) test. This test results in identical conclusions as the binomial test in which the normal approximation is used as described above. The chi-square distribution is frequently used in statistical tests involving counts and proportions, as discussed in chapter 15. Here, we will show the application to fourfold tables (2 × 2 tables), the comparison of proportions in two independent groups.

The chi-square distribution is appropriate where the normal approximation to the distribution of discrete variables can be applied. In particular, when comparing two proportions, the chi-square distribution with 1 d.f. can be used to approximate probabilities. (The values for the

** The continuity correction can make a difference when making decisions based on the α level, when the statistical test is "just significant" (e.g., $p = 0.04$ for a test at the 5% level). The correction makes the test "less significant."

Table 5.13 Result of the Experiment Shown in Table 5.12 in the Form of a Fourfold Table

	Group		
	I	II	Total
Number of successes	16	18	34
Number of failures	4	12	16
Total	20	30	50

χ^2 distribution with one d.f. are exactly the square of the corresponding normal deviates. For example, the "95%" cutoff point for the chi-square distribution with 1 d.f. is 3.84, equal to 1.96^2.)

The use of the chi-square distribution to test for differences of proportions in two groups has two advantages: (a) the computations are easy and (b) a continuity correction can be easily applied. The reader may have noted that a continuity correction was not used in the examples for the comparison of two independent groups described above. The correction was not included because the computation of the correction is somewhat complicated. In the chi-square test, however, the continuity correction is relatively simple. The correction is most easily described in the context of an example. We will demonstrate the chi-square test using the data in Table 5.12. We can think of these data as resulting from a clinical trial where groups I and II represent two comparative drugs. The same results are presented in the *fourfold table* shown in Table 5.13.

The chi-square statistic is calculated as follows:

$$\chi^2 = \sum \frac{(O - E)^2}{E}, \tag{5.19}$$

where O is the observed number in a cell (there are four cells in the experiment in Table 5.13; a cell is the intersection of a row and column; the upper left-hand cell, number of successes in group I, has the value 16 contained in it), and E the expected number in a cell.

The expected number is the number that would result if each group had the same proportion of successes and failures. The best estimate of the common p (proportion of successes) is the pooled value, as calculated in the test using the normal approximation above [Eq. (5.17)]. The pooled, p, $\hat{p}_0$, is 0.68 (34/50). With a probability of success of 0.68 (34/50), we would expect "13.6" successes for group I (20×0.68). The expected number of failures is $20 \times 0.32 = 6.4$. The expected number of failures can also be obtained by subtracting 13.6 from the total number of observations in group I, $20 - 13.6 = 6.4$. Similarly, the expected number of successes in group II is $30 \times 0.68 = 20.4$. Again the number, 20.4, could have been obtained by subtracting 13.6 from 34.

This concept (and calculation) is illustrated in Table 5.14, which shows the expected values for Table 5.13. The marginal totals (34, 16, 20, and 30) in the "expected value" table are the same as in the original table, Table 5.13. In order to calculate the expected values, multiply the two marginal totals for a cell and divide this value by the grand total. This simple way of calculating

Table 5.14 Expected Values for the Experiment Shown in Table 5.13

	Group		
	I	II	Total
Expected number of successes	13.6	20.4	34
Expected number of failures	6.4	9.6	16
Total	20.0	30.0	50

the expected values will be demonstrated for the upper left-hand cell, where the observed value is 16. The expected value is

$$\frac{(20)(34)}{50} = 13.6.$$

Once the expected value for one cell is calculated, the expected values for the remaining cells can be obtained by subtraction.

Expected successes in group II $= 34 - 13.6 = 20.4$.
Expected failures in group I $= 20 - 13.6 = 6.4$.
Expected failures in group II $= 16 - 6.4 = 9.6$.

Given the marginal totals and the value for any *one* cell, the values for the other three cells can be calculated. Once the expected values have been calculated, the chi-square statistic is evaluated according to Eq. (5.19).

$$\sum \frac{(O-E)^2}{E} = \frac{(16-13.6)^2}{13.6} + \frac{(18-20.4)^2}{20.4} + \frac{(4-6.4)^2}{6.4} + \frac{(12-9.6)^2}{9.6} = 2.206$$

The numerator of each term is $(\pm 2.4)^2 = 5.76$. Therefore, the computation of χ^2 can be simplified as follows:

$$\chi^2 = (O-E)^2 \left(\frac{1}{E_1} + \frac{1}{E_2} + \frac{1}{E_3} + \frac{1}{E_4} \right), \tag{5.20}$$

where E_1 through E_4 are the expected values for each of the four cells.

$$\chi^2 = (2.4)^2 \left(\frac{1}{13.6} + \frac{1}{20.4} + \frac{1}{6.4} + \frac{1}{9.6} \right) = 2.206$$

One can show that this computation is exactly equal to the square of the Z value using the normal approximation to the binomial. (See Exercise Problem 11.)

The d.f. for the test described above (the fourfold table) are equal to 1. In general, the d.f. for an $R \times C$ contingency table, where R is the number of rows and C is the number of columns, are equal to $(R-1)(C-1)$. The analysis of $R \times C$ tables is discussed in chapter 15.

Table IV.5, a table of points in the cumulative chi-square distribution, shows that a value of 3.84 is needed for significance at the 5% level (1 d.f.). Therefore, the test in this example is not significant; that is, the proportion of successes in group I is not significantly different from that in group II, 0.8 and 0.6, respectively.

To illustrate further the computations of the chi-square statistic and the application of the continuity correction, we will analyze the data in Example 10, where the normal approximation to the binomial was used for the statistical test. Table 5.15 shows the observed and expected values for the results of this preclinical study.

Table 5.15 Observed and Expected Values for Preclinical Carcinogenicity Study[a]

	Drug	Placebo	Total
Animals with carcinoma	9(7.2)	6 (7.8)	15
Animals without carcinoma	51(52.8)	59 (57.2)	110
Total	60	65	125

[a]Parenthetical values are expected values.

The uncorrected chi-square analysis results in a value of 0.98, $(0.99)^2$. (See Exercise Problem 18.) The continuity correction is applied using the following rule: If the fractional part of the difference $(O - E)$ *is* larger than 0 but ≤ 0.5, delete the fractional part. If the fractional part is greater than 0.5 or exactly 0, "reduce the fractional part to 0.5." Some examples should make the application of this rule clearer.

$O - E$	Corrected for continuity
3.0	2.5
3.2	3.0
3.5	3.0
3.9	3.5
3.99	3.5
4.0	3.5

In the example above, $O - E = \pm 1.8$. Therefore, correct this value to ± 1.5. The corrected chi-square statistic is [Eq. (5.20)]

$$(1.5)^2 \left(\frac{1}{7.2} + \frac{1}{7.8} + \frac{1}{52.8} + \frac{1}{57.2} \right) = 0.68.$$

In this example, the result is not significant using either the corrected or uncorrected values. However, when chi-square is close to significance at the α level, the continuity correction can make a difference. The continuity correction is more apparent in its effect on the computation of chi-square in small samples. With large samples, the correction makes less of a difference.

The chi-square test, like the normal approximation, is an approximate test, applying a continuous distribution to discrete data. The test is valid (close to correct probabilities) when the expected value in each cell is at least 5. This is an approximate rule. Because the rule is conservative, in some cases, an expected value in one or more cells of less than 5 can be tolerated. However, one should be cautious in applying this test if the expected values are too small.

5.2.6 Confidence Intervals for Proportions

Examples of the formation of a confidence interval for a proportion have been presented earlier in this chapter (Example 3). Although the confidence interval for the binomial is calculated using the standard deviation of the binomial based on the sample proportion, we should understand that in most cases, the s.d. is unknown. The sample standard deviation is an estimate of the true s.d., which for the binomial depends on the true value of the proportion or probability. However, when we use the sample estimate of the s.d. for the calculations, the confidence interval and statistical tests are valid using criteria based on the normal distribution (Table IV.2). We do not use the *t* distribution as in the procedures discussed previously.

The confidence interval for the true proportion or binomial probability, p_0 is

$$\hat{p} \pm Z\sqrt{\frac{\hat{p}\hat{q}}{N}}, \tag{5.3}$$

where $\hat{p}$ is the observed proportion in a sample of size N. The value of Z depends on the confidence coefficient (e.g., 1.96 for a 95% interval). Of 500 tablets inspected, 20 were found to be defective ($\hat{p} = 20/500 = 0.04$). A 95% confidence interval for the true proportion of defective tablets is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{N}} = 0.04 \pm 1.96\sqrt{\frac{(0.04)(0.96)}{500}}$$
$$= 0.04 \pm 0.017 = 0.023 \text{ to } 0.057.$$

To obtain a *confidence interval for the difference of two proportions* (two independent groups), when the "underlying proportions, p_1 and p_2 are not hypothesized to be equal (7), use the following formula:

$$(\hat{p}_1 - \hat{p}_2) \pm Z\sqrt{\frac{\hat{p}_1\hat{q}_1}{N_1} + \frac{\hat{p}_2\hat{q}_2}{N_2}} \tag{5.21}$$

where $\hat{p}_1$ and $\hat{p}_2$ are the observed proportions in groups I and II, respectively, and N_1, and N_2 are the respective sample sizes of the two groups. Z is the appropriate normal deviate (1.96 for a 95% confidence interval).

In the example of incidence of headaches in two groups of patients, the proportion of headaches observed in group I was $35/212 = 0.165$ and the proportion in group II was $46/196 = 0.235$. A 95% confidence interval for the difference of the two proportions, calculated from Eq. (5.21), is

$$(0.235 - 0.165) \pm 1.96\sqrt{\frac{(0.165)(0.835)}{212} + \frac{(0.235)(0.765)}{196}}$$

$$= 0.07 \pm 0.078 = -0.008 \text{ to } 0.148.$$

The difference between the two proportions was not significant at the 5% level in a two-sided test (see "Test for Comparison of Proportions from Two Independent Groups" in sect. 5.2.4). Note that 95% confidence interval covers 0, the difference specified in the null hypothesis ($H_0 : p_1 - p_2 = 0$).††

Fleiss [7] and Hauck and Anderson [8] recommend the use of a continuity correction for the construction of confidence intervals that gives better results than that obtained without a correction [Eq. (5.21)]. If a 90% or 95% interval is used, the Yates correction works well if N_1p_1, N_1q_1, N_2p_2, and N_2q_2 are all greater than or equal to 3. The 99% interval is good for N_1p_1, N_1q_1, N_2p_2, and N_2q_2 all greater than or equal to 5. The correction is $1/2N_1 + 1/2N_2$. Applying the correction to the previous example, a 95% confidence interval is

$$\begin{aligned}&(0.235 - 0.165)\\ &\quad \pm \{1.96\,[(0.165)(0.835)/212 + (0.235)(0.765)/196]^{1/2} + (1/424 + 1/392)\}\\ &\quad = 0.070 \pm 0.0825 = -0.0125 \text{ to } 0.1525.\end{aligned}$$

We have noted previously that if the hypothesis test of equality of two proportions is statistically significant, the confidence interval for the difference of the proportions will not cover zero (and vice versa). Sometimes, this does not hold when comparing two proportions because the formulation for the hypothesis test of equal proportions is different from the confidence interval calculation. In this case, Fleiss [7] recommends changing the hypothesis test statistic by replacing the denominator in equation 5.17 with the square root of $(p_1q_1)/N_1 + (p_2q_2)/N_2$.

An approach to sample size requirements using confidence intervals for bioequivalence trials with a binomial variable is given in section 11.4.8.

5.3 COMPARISON OF VARIANCES IN INDEPENDENT SAMPLES

Most of the statistical tests presented in this book are concerned with means. However, situations arise where variability is important as a measure of a process or product performance. For example, when mixing powders for tablet granulations, one may be interested in measuring the homogeneity of the mix as may be indicated in validation procedures. The "degree" of homogeneity can be determined by assaying different portions of the mix, and calculating the

†† The form of the confidence interval [Eq. (5.21)] differs from the form of the statistical test in that the latter uses the pooled variance [Eq. (5.17)]. Therefore, this relationship will not always hold for the comparison of two proportions.

standard deviation or variance. (Sample weights equal to that in the final dosage form are most convenient.) A small variance would be associated with a relatively homogeneous mix, and vice versa. Variability is also often of interest when assaying drug blood levels in a bioavailability study or when determining a clinical response to drug therapy. We will describe statistical tests appropriate for two situations: the comparison of two variances from independent samples, and the comparison of variances in related (paired) samples. The test for related samples will be presented in chapter 7 because methods of calculation involve material presented there. The test for the comparison of variances in independent samples described here assumes that the data in each sample are independent and normally distributed.

The notion of significance tests for two variances is similar to the tests for means (e.g., the *t* test). The null hypothesis is usually of the form

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

For a two-sided test, the alternative hypothesis admits the possibility of either variance being larger or smaller than the other

$$H_0 : \sigma_1^2 \neq \sigma_2^2.$$

The statistical test consists of calculating the ratio of the two sample variances. The ratio has an *F* distribution with $(N_1 - 1)$ d.f. in the numerator and $(N_2 - 1)$ d.f. in the denominator. To determine if the ratio is "significant" (i.e., the variances differ), the observed ratio is compared to appropriate table values of *F* at the α level. The *F* distribution is not symmetrical and, in general, to make statistical decisions, we would need *F* tables with both upper and lower cutoff points.

Referring to Figure 5.12, if the *F* ratio falls between F_L and F_U the test is not significant. We do not reject the null hypothesis of equal variances. If the *F* ratio is below F_L or above F_U, we reject the null hypothesis and conclude that the variances differ (at the 5% level, the shaded area in the example of Fig. 5.12). The *F* table to test the equality of two variances is the same as that used to determine significance in analysis of variance tests to be presented in chapter 8 (Table IV.6). However, *F* tables for ANOVA usually give only the upper cutoff points (F_U,0.05 in Fig. 5.12, for example).

Nevertheless, it is possible to perform a two-sided test for two variances using the one-tailed *F* table (Table IV.6) by forming the ratio with the *larger variance in the numerator.* Thus, the ratio will always be equal to or greater than 1. The ratio is then referred to the usual ANOVA *F* table, but the level of significance is twice that stated in the table. For example, the values that must be exceeded for significance in Table IV.6 represent cutoff points at the 20%, 10% or 2% level if the larger variance is in the numerator. For significance at the 5% level, use Table 5.16, a brief table of the upper 0.025 cutoff points for some *F* distributions.

To summarize, for a two-sided test at the 5% level, calculate the ratio of the comparative variances with the larger variance in the numerator. (Clearly, if the variances in the two groups are identical, there is no need to perform a test of significance.) To be significant at the 5% level, the ratio must be equal to or greater than the tabulated upper 2.5% cutoff points (Table 5.16). For significance at the 10% level or 20% level, for a two-sided test, use the upper 5% or 10% points in Table IV.6A.

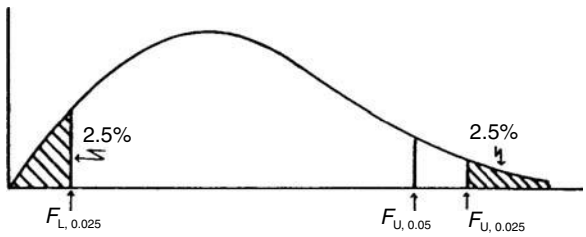


Figure 5.12 Example of two-sided cutoff points in an *F* distribution.

Table 5.16 Brief Table of Upper 0.025 Cutoff Points of the *F* Distribution

Degrees of freedom in denominator	Degrees of freedom in numerator											
	2	**3**	**4**	**5**	**6**	**8**	**10**	**15**	**20**	**25**	**30**	∞
2	39.0	39.2	39.3	39.3	39.3	39.4	39.4	39.4	39.5	39.5	39.5	39.5
3	16.0	15.4	15.1	14.9	14.7	14.5	14.4	14.3	14.2	14.1	14.1	13.9
4	10.6	10.0	9.6	9.4	9.2	9.0	8.8	8.7	8.6	8.5	8.5	8.3
5	8.4	7.8	7.4	7.2	7.0	6.8	6.6	6.4	6.3	6.3	6.2	6.0
6	7.3	6.6	6.2	6.0	5.8	5.6	5.5	5.3	5.2	5.1	5.1	4.9
7	6.5	5.9	5.5	5.3	5.1	4.9	4.8	4.6	4.5	4.4	4.4	4.1
8	6.1	5.4	5.1	4.8	4.7	4.4	4.3	4.1	4.0	3.9	3.9	3.7
9	5.7	5.1	4.7	4.5	4.3	4.1	4.0	3.8	3.7	3.6	3.6	3.3
10	5.5	4.8	4.5	4.2	4.1	3.9	3.7	3.5	3.4	3.4	3.3	3.1
15	4.8	4.2	3.8	3.6	3.4	3.2	3.1	2.9	2.8	2.7	2.6	2.4
20	4.5	3.9	3.5	3.3	3.1	2.9	2.8	2.6	2.5	2.4	2.4	2.1
24	4.3	3.7	3.4	3.2	3.0	2.8	2.6	2.4	2.3	2.3	2.2	1.9
30	4.2	3.6	3.3	3.0	2.9	2.7	2.5	2.3	2.2	2.1	2.1	1.8
40	4.1	3.5	3.1	2.9	2.7	2.5	2.4	2.2	2.1	2.0	1.9	1.6
∞	3.7	3.1	2.8	2.6	2.4	2.2	2.1	1.8	1.7	1.6	1.6	1.0

For a *one-sided test*, if the null hypothesis is

$$H_0\colon \sigma_A^2 \geq \sigma_B^2 \qquad H_0\colon \sigma_A^2 < \sigma_B^2.$$

Perform the test only if S_A^2 is smaller than S_B^2, with S_B^2 in the numerator. (If S_A^2 is equal to or greater than S_B^2, we cannot reject the null hypothesis.) Refer the ratio to Table IV.6 for significance at the 5% (or 1%) level. (The test is one -sided.)

One should appreciate that this statistical test is particularly sensitive to departures from the assumptions of normality and independence of the two comparative groups.

An example should clarify the procedure. Two granulations were prepared by different procedures. Seven random samples of powdered mix of equal weight (equal to the weight of the final dosage form) were collected from each batch and assayed for active material, with the results shown in Table 5.17. The test is to be performed at the 5% level: $H_0 : \sigma_1^2 = \sigma_2^2$; $H_a\colon \sigma_1^2 \neq \sigma_2^2$. For a two-sided test, we form the ratio of the variances with a σ_B^2, the larger variance in the numerator.

$$F = \frac{1.297}{0.156} = 8.3.$$

The tabulated F value with 6 d.f. in the numerator and denominator (Table 5.16) is 5.8. Therefore, the variances can be considered significantly different ($P < 0.05$); granulation B is more variable than granulation A. If the test were performed at the 10% level, we would refer to the upper 5% points in Table IV.6, where a value greater than 4.28 would be significant.

Table 5.17 Assays from Samples from Two Granulations

Granulation *A*		**Granulation *B***	
20.6	20.7	20.2	19.0
20.9	19.8	21.5	21.8
20.6	20.4	18.9	20.4
	21.0		21.0
$\overline{X} = 20.57$	$S^2 = 0.156$	$\overline{X} = 20.4$	$S^2 = 1.297$

If the test were *one sided, at the* 5% *level*, for example, with the null hypothesis

$$H_0 : \sigma_A^2 \geq \sigma_B^2 \qquad H_a : \sigma_A^2 < \sigma_B^2,$$

the ratio $1.297/0.156 = 8.3$ would be referred to Table IV.6 for significance. Now, a value greater than 4.28 would be significant at the 5% level.

If more than two variances are to be compared, the F test discussed above is not appropriate. Bartlett's test is the procedure commonly used to test the equality of more than two variances [1], as described in the following paragraph.

5.4 TEST OF EQUALITY OF MORE THAN TWO VARIANCES

The test statistic computation is shown in Eq. (5.22)

$$\chi^2 = \sum (N_i - 1) \ln S^2 - \sum [(N_i - 1) \ln S_1^2], \tag{5.22}$$

where S^2 is the pooled variance and S_i^2 is the variance of the ith sample.

The computations are demonstrated for the data of Table 5.18. In this example, samples of a granulation were taken at four different locations in a mixer. Three samples were analyzed in each of three of the locations, and five samples analyzed in the 4th location. The purpose of this experiment was to test the homogeneity of the mix in a validation experiment. Part of the statistical analysis requires an estimate of the variability within each location. The statistical test (analysis of variance, chap. 8) assumes homogeneity of variance within the different locations. Bartlett's test allows us to test for the homogeneity of variance (Table 5.8).

The pooled variance is calculated as the weighted average of the variances, where the weights are the d.f. $(N_i - 1)$.

$$\text{Pooled } S^2 = \frac{2 \times 3.6 + 2 \times 4.7 + 2 \times 2.9 + 4 \times 8.3}{2 + 2 + 2 + 4} = 5.56$$

$$\sum (N_i - 1) = 10$$

$$\sum [(N_i - 1) \ln S_i^2 = 2(1.2809) + 2(1.5476) + 2(1.0647)] + 4(2.1163) = 16.2516$$

$$\chi^2 = 10 \times \ln (5.56) - 16.2516 = 0.904.$$

To test χ^2 for significance, compare the result to the tabulated value of χ^2 (Table IV.5) with 3 d.f. (1 less than the number of variances being compared) at the appropriate significance level. A value of 7.81 is needed for significance at the 5% level. Therefore, we conclude that the variances do not differ. A significant value of χ^2 means that the variances are not all equal. This test is very sensitive to non-normality. That is, if the variances come from non-normal populations, the conclusions of the test may be erroneous.

See Exercise Problem 22 for another example where Bartlett's test can be used to test the homogeneity of variances.

Table 5.18 Results of Variability of Assays of Granulation at Six Locations in a Mixer

Location	*N*	*N-1*	Variance (S^2)	ln(S^2)
A	3	2	3.6	1.2809
B	3	2	4.7	1.5476
C	3	2	2.9	1.0647
D	5	4	8.3	2.1163

Table 5.19 Short Table of Lower and Upper Cutoff Points for Chi-Square Distribution

Degrees of freedom	Lower 2.5%	Lower 5%	Upper 95%	Upper 97.5%
2	0.0506	0.1026	5.99	7.38
3	0.216	0.352	7.81	9.35
4	0.484	0.711	9.49	11.14
5	0.831	1.15	11.07	12.83
6	1.24	1.64	12.59	14.45
7	1.69	2.17	14.07	16.01
8	2.18	2.73	15.51	17.53
9	2.70	3.33	16.92	19.02
10	3.25	3.94	18.31	20.48
15	6.26	7.26	25.00	27.49
20	9.59	10.85	31.41	34.17
30	16.79	18.49	43.77	46.98
60	40.48	43.19	79.08	83.30
120	91.58	95.76	146.57	152.21

5.5 CONFIDENCE LIMITS FOR A VARIANCE

Given a sample variance, a confidence interval for the variance can be constructed in a manner similar to that for means. $S^2/\sigma^2$is distributed as $\chi^2/\text{d.f.}$ The confidence interval can be obtained from the chi-square distribution, using the relationship shown in Eq. (5.23).

$$\frac{S^2(n-1)}{\text{chi-square}_{\alpha/2}} \geq \sigma^2 \geq \frac{S^2(n-1)}{\text{chi-square}_{1-\alpha/2}} \tag{5.23}$$

For example, a variance estimate based on 10 observations is 4.2, with 9 d.f. For a 90% two-sided confidence interval, we put 5% of the probability in each of the lower and upper tails of the χ^2 distribution. From Table 5.19 and Eq. (5.23), the upper limit is

$$S^2(9/3.33) = 4.2(9/3.33) = 11.45.$$

The lower limit is

$$4.2(9/16.92) = 2.23.$$

The values, 3.33 and 16.92, are the cutoff points for 5% and 95% of the chi-square distribution with 9 d.f.. Thus, we can say that with 90% probability, the true variance is between 2.23 and 11.45. Exercise Problem 23 shows an example of a one-sided confidence interval for the variance from a content uniformity test.

5.5.1 Rationale for USP Content Uniformity Test

The USP content uniformity test was based on the desire for a plan that would limit acceptance to lots with sigma (RSD) less than 10% [9]. The main concern is to prevent the release of batches of product with excessive units outside of 75% to 125% of the labeled dose that may occur for lots with a large variability. If the observed RSD for 10 units is less than 6%, one can demonstrate that there is less than 0.05 probability that the true RSD of the lot is greater than 10%. A two-sided 90% confidence interval for an RSD of 6 for $N = 10$, can be calculated by taking the square root of the interval for the variance. In this example, the variance is 36 (RSD = 6). Following the logic of the previous example, the upper limit of the 90% confidence interval for the variance is $6^2(9/3.33) = 97.3$. Since the upper limit represents a one-sided 95% confidence limit, the upper limit for the standard deviation(s) is $\sqrt{97.3}$, approximately 10. See also Exercise Problem 24 at the end of this chapter.

5.6 TOLERANCE INTERVALS

Tolerance intervals have a wide variety of potential applications in pharmaceutical and clinical data analysis. A tolerance interval describes an interval in which a given percentage of the individual items lie, with a specified probability. This may be expressed as

Probability($L \leq$ % of population $\leq U$) where L is the lower limit and U is the upper limit.

For example, a tolerance interval might take the form of a statement such as, "There is 99% probability that 95% of the population is between 85 and 115." More specifically, we might say that there is 99% probability that 95% of the tablets in a batch have a potency between 85% and 115%. In order to be able to compute tolerance intervals, we must make an assumption about the data distribution. As is typical in statistical applications, the data will be assumed to have a normal distribution. In order to compute the tolerance interval, we need an estimate of the mean and standard deviation. These estimates are usually taken from a set of observed experimental data.

Given the d.f. for the estimated s.d., the limits can be computed from Table IV.19 in appendix IV. The factors in Table IV.19 represent multipliers of the standard deviation, similar to a confidence interval. Therefore, using these factors, the tolerance interval computation is identical to the calculation of a confidence interval.

$$P\% \text{ tolerance interval containing } X\% \text{ of the population} = \overline{X} \pm t' \text{ (s.d.)}$$

where t' is the appropriate factor found in Table IV.19.

The following examples are intended to make the calculation and interpretation clearer.

Table 5.20 Summary of Tests

Test		Section
Mean of single population	$t = \dfrac{\overline{X} - \mu}{S\sqrt{1/N}}$	5.2.1
Comparison of means from two independent populations (variances known)	$Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}}$	5.2.2
Comparisons of means from two independent populations (variance unknown)	$t = \dfrac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{1/N_1 + 1/N_2}}$	5.2.2
Comparison of means from two related samples (variance unknown)[a]	$t = \dfrac{\delta}{S\sqrt{1/N}}$	5.2.3
Proportion from a single population[b]	$Z = \dfrac{p - p_0}{\sqrt{p_0 q_0 N)}}$	
Comparison of two proportions from independent groups[b]	$Z = \dfrac{p - p_0}{\sqrt{p_0 q_0 (1/N_1 + 1/N_2)}}$	5.2.4
Comparison of variances (two-sided test)	$F = \dfrac{S_1^2}{S_2^2}$ $(S_1^2 > S_2^2)$	5.2.4
Confidence limits for variance	$\dfrac{(n^2-1)S^2_{(n-1)}}{\chi^2_{\alpha/2}} \geq \sigma^2 \geq \dfrac{(n-1)S^2_{(n-1)}}{\chi^2_{1-\alpha/2}}$	5.5

[a] If the variance is known, use the normal distribution.
[b] A continuity correction may be used (5.16 and 5.20).

Example 1. A batch of tablets was tested for content uniformity. The mean of the 10 tablets tested was 99.1% and the s.d. was 2.6%. Entering Table IV.19, for a 99% tolerance interval that contains 99.9% of the population with $N = 10$, the factor, $t\prime = 7.129$. Assuming a normal distribution of tablet potencies, we can say with 99% probability (99% "confidence") that 99.9% of the tablets are within $99.1\ \% \pm 7.129 \times 2.6 = 99.1\ \% \pm 18.5 = 80.6\ \%$ to 117.6 %.

Example 2. In a bioequivalence study using a crossover design with 24 subjects, the ratio of test product to standard product was computed for each subject. One of the proposals for assessing individual bioequivalence is to compute a tolerance interval to estimate an interval that will encompass a substantial proportion of subjects who take the drug. The average of the 24 ratios was 1.05 with a s.d. of 0.3. A tolerance interval is calculated that has 95% probability of containing 75% of the population. The factor from Table IV.19 for $N = 24$ and 95% confidence is 1.557. The tolerance interval is $1.05 \pm 1.557 \times 0.3 = 1.05 \pm 0.47$. Thus, we can say that 75% of the patients will have a ratio between 0.58 and 1.52 with 95% probability. One of the problems with such an approach to individual equivalence is that the interval is dependent on the variability, and highly variable drugs will always show a wide variation of the ratio for different products. Therefore, using this interval as an acceptance criterion for individual equivalence may not be very meaningful. Also, this computation assumes a normal distribution, and individual ratios may deviate significantly from a normal distribution.

Table 5.20 summarizes some tests discussed in this chapter.

KEY TERMS

Alpha level
Alternative hypothesis
Bartlett's test
Behrens–Fisher test
Beta error
Bias
Binomial trials
Blinding
Cells
Chi-square test
Confidence interval
Continuity correction
Critical region
Crossover design
Cumulative normal distribution
Degrees of freedom
Delta
Error
Error of first kind
Estimation
Expected values
Experimental error
Fourfold table
F test
Hypothesis testing
Independence
Independent groups
Level of significance
Marginal totals
Nonsignificance
Normal curve test
Null hypothesis
One-sample test
One-sided test
One-way analysis of variance
Paired-sample *t* test
Parallel-groups design
Parameters
Pooled proportion
Pooled variance
Power
Preference tests
Randomization
Region of rejection
Sample size
Sensitive
Significance
t distribution
t test
Tolerance interval
Two-by-two table
Two independent groups *t* test
Two-tailed (sided) test
Uncontrolled study
Variance
Yates correction
Z transformation

EXERCISES

1. Calculate the probability of finding a value of 49.8 or less if $\mu = 54.7$ and $\sigma = 2$.

2. If the variance of the population of tablets in Table 5.1 were known to be 4.84, compute a 99% confidence interval for the mean.

3. (a) Six analysts perform an assay on a portion of the same homogeneous material with the following results: 5.8, 6.0, 5.7, 6.1, 6.0, and 6.1. Place 95% confidence limits on the true mean.
   (b) A sample of 500 tablets shows 12 to be defective. Place a 95% confidence interval on the percent defective in the lot.
   (c) Place a 95% confidence interval on the difference between two products in which 50 of 60 patients responded to product *A*, and 25 of 50 patients responded to product *B*.

4. (a) Quality control records show the average tablet weight to be 502 mg with a standard deviation of 5.3. There are sufficient data so that these values may be considered known parameter values. A new batch shows the following weights from a random sample of six tablets: 500, 499, 504, 493, 497, and 495 mg. Do you believe that the new batch has a different mean from the process average?
   (b) Two batches of tablets were prepared by two different processes. The potency determinations made on five tablets from each batch were as follows: batch *A*: 5.1, 4.9, 4.6, 5.3, 5.5; batch *B*: 4.8, 4.8, 5.2, 5.0, 4.5. Test to see if the means of the two batches are equal.
   (c) Answer part (a) if the variance were unknown. Place a 95% confidence interval on the true average weight.

5. (a) In part (b) of Problem 4, calculate the variance and the standard deviation of the 10 values as if they were one sample. Are the values of the s.d. and S^2 smaller or larger than the values calculated from "pooling"?
   (b) Calculate the pooled s.d. above by "averaging" the s.d.'s from the two samples. Is the result different from the "pooled" s.d. as described in the text?

6.

Batch 1 (drug)	Pass/fail (improve, worsen)	Batch 2 (placebo)	Pass/fail (improve, worsen)
10.1	P	9.5	F
9.7	F	8.9	F
10.1	P	9.4	F
10.5	P	10.4	P
12.3	P	9.9	F
11.8	P	10.1	P
9.6	F	9.0	F
10.0	F	9.7	F
11.2	P	9.9	F
11.3	P	9.8	F

   (a) What are the mean and s.d. of each batch? Test for difference between the two batches using a *t* test.
   (b) What might be the "population" corresponding to this sample? Do you think that the sample size is large enough? Why? Ten objects were selected from each batch for this test. Is this a good design for comparing the average results from two batches?
   (c) Consider values above 10.0 a success and values 10.00 or less a failure. What is the proportion of successes for batch 1 and batch 2? Is the proportion of successes in batch 1 different from the proportion in batch 2 (5% level)?
   (d) Put 95% confidence limits on the proportion of successes with all data combined.

7. A new analytical method is to be compared to an old method. The experiment is performed by a single analyst. She selects four batches of product at random and obtains the following results.

Batch	Method 1	Method 2
1	4.81	4.93
2	5.44	5.43
3	4.25	4.30
4	4.35	4.47

(a) Do you think that the two methods give different results on the average?
(b) Place 95% confidence limits on the true difference of the methods.

8. The following data for blood protein (g/100 mL) were observed for the comparison of two drugs. Both drugs were tested on each person in random order.

Patient	Drug *A*	Drug *B*
1	8.1	9.0
2	9.4	9.9
3	7.2	8.0
4	6.3	6.0
5	6.6	7.9
6	9.3	9.0
7	7.6	7.9
8	8.1	8.3
9	8.6	8.2
10	8.3	8.9
11	7.0	8.3
12	7.7	8.8

(a) Perform a statistical test for drug differences at the 5% level.
(b) Place 95% confidence limits on the average differences between drugs A and B.

9. For examples 10 and 11, calculate the pooled p and q (p_0 and q_0).

10. In Example 4, perform a t test if the mean were 16.7 instead of 15.9.

11. Use the normal approximation and chi-square test (with and without continuity correction) to answer the following problem. A placebo treatment results in 8 patients out of 100 having elevated blood urea nitrogen (BUN) values. The drug treatment results in 16 of 100 patients having elevated values. Is this significantly different from the placebo?

12. Quality control records show that the average defect rate for a product is 2.8%. Two hundred items are inspected and 5% are found to be defective in a new batch. Should the batch be rejected? What would you do if you were the director of quality control? Place confidence limits on the *percent* defective and the *number* defective (out of 200).

¶¶**13. In a batch size of 1,000,000,5000 tablets are inspected and 50 are found defective.
(a) Put 95% confidence limits on the true number of defectives in the batch.
(b) At $\alpha = 0.05$, do you think that there could be more than 2% defective in the batch?
(**c) If you wanted to estimate the true proportion of defectives within $\pm$ 0.1% with 95% confidence, how many tablets would you inspect?

14. In a clinical test, 60 people received a new drug and 50 people received a placebo. Of the people on the new drug, 40 of the 60 showed a positive response and 25 of the 50 people on placebo showed a positive response. Perform a statistical test to determine if

¶¶The double asterisk indicates optional, more difficult problems.

the new drug shows more of an effect than the placebo. Place a 95% confidence interval on the difference of proportion of positive response in the two test groups.

15. In a paired preference test, each of 100 subjects was asked to choose the preference between A and B. Of these 100, 60 showed no preference, 30 preferred A, and 10 preferred B. Is A significantly preferred to B?

16. Over a long period of time, a screening test has shown a response rate for a control of 20%. A new chemical shows 9 positive results in 20 observations (45%). Would you say that this candidate is better than the control? Place 99% confidence limits on the true response rate for the new chemical.

17. Use the chi-square test with the continuity correction to see if there is a significant difference in the following comparison. Two batches of tablets were made using different excipients. In batch A, 10 of 100 tablets sampled were chipped. In batch B, 17 of 95 tablets were chipped. Compare the two batches with respect to proportion chipped at the 5% level.

18. Show that the uncorrected value of chi-square for the data in Table 5.15 is 0.98.

19. Use the chi-square test, with continuity correction, to test for significance (5% level) for the data in Example 11.

20. Perform a statistical test to compare the variances in the two groups in Problem 6. $H_0 : \sigma_1^2 = \sigma_2^2; \quad H_a : \sigma_1^2 \neq \sigma_2^2$. Perform the test at the 10% level.

21. Compute the value of the corrected χ^2 statistic for data of Example 11 in 5.2.4. Compute the t value as recommended by D'Agostino et al. Compare the uncorrected value of Z with these results.

22. The homogeneity of a sample taken from a mixer was tested after 5, 10, and 15 minutes. The variances of six samples taken at each time were 16.21, 1.98, and 2.02. Based on the results of Bartlett's test for homogeneity of variances, what are your conclusions?

23. Six blend samples (unit dose size) show a variance of 9% (RSD = 3%). Compute a 95% one-sided upper confidence interval for the variance. Is this interval too large based on the official limit of 6% for RSD?

24. The USP content uniformity test for 30 units states that the RSD should not exceed 7.8%. Show that there is a 5% probability that the true RSD is less than 10%.

REFERENCES

1. Westlake WJ. Symmetrical confidence intervals for bioequivalence trials. Biometrics 1976; 32:741–744.
2. Westlake WJ. Bioavailability and bioequivalence of pharmaceutical formulations. In: Peace KE, ed. Statistical Issues in Drug Research and Development. New York: Marcel Dekker, 1990.
3. Snedecor GW, Cochran WG. Statistical Methods, 6th ed. Ames, IA: Iowa State University Press, 1967.
4. Cochran WG. Sampling Techniques, 3rd ed. New York: Wiley, 1967.
5. Snedecor GW, Cochran WG. Statistical Methods, 8th ed. Ames, IA: Iowa State University Press, 1989.
6. D'Agostino RB, Chase W, Belanger A. The appropriateness of some common procedures for testing the equality of two independent binomial populations. Am Stat 1988; 42:198
7. Fleiss J. Statistical Methods for Rates and Proportions, 2nd ed. New York: Wiley, 1981.
8. Hauck W, Anderson S. A comparison of large sample CI methods for the difference of two binomial probabilities. Am Stat 1986; 40:318.
9. Cowdery S, Michaels T. Pharmacopeial Forum. 1980:614.

6 Sample Size and Power

The question of the size of the sample, the number of observations, to be used in scientific experiments is of extreme importance. Most experiments beg the question of sample size. Particularly when time and cost are critical factors, one wishes to use the minimum sample size to achieve the experimental objectives. Even when time and cost are less crucial, the scientist wishes to have some idea of the number of observations needed to yield sufficient data to answer the objectives. An elegant experiment will make the most of the resources available, resulting in a sufficient amount of information from a minimum sample size. For simple comparative experiments, where one or two groups are involved, the calculation of sample size is relatively simple. A knowledge of the α level (level of significance), β level ($1 -$ power), the standard deviation, and a meaningful "practically significant" difference is necessary in order to calculate the sample size.

Power is defined as $1 - \beta$ (i.e., $\beta = 1 -$ power). Power is the ability of a statistical test to show significance if a specified difference truly exists. The magnitude of power depends on the level of significance, the standard deviation, and the sample size. Thus power and sample size are related.

In this chapter, we present methods for computing the sample size for relatively simple situations for normally distributed and binomial data. The concept and calculation of power are also introduced.

6.1 INTRODUCTION

The question of sample size is a major consideration in the planning of experiments, but may not be answered easily from a scientific point of view. In some situations, the choice of sample size is limited. Sample size may be dictated by official specifications, regulations, cost constraints, and/or the availability of sampling units such as patients, manufactured items, animals, and so on. The USP content uniformity test is an example of a test in which the sample size is fixed and specified [1].

The sample size is also specified in certain quality control sampling plans such as those described in MIL-STD-105E [2]. These sampling plans are used when sampling products for inspection for attributes such as product defects, missing labels, specks in tablets, or ampul leakage. The properties of these plans have been thoroughly investigated and defined as described in the document cited above. The properties of the plans include the chances (probability) of rejecting or accepting batches with a known proportion of rejects in the batch (sect. 12.3).

Sample-size determination in comparative clinical trials is a factor of major importance. Since very large experiments will detect very small, perhaps clinically insignificant, differences as being statistically significant, and small experiments will often find large, clinically significant differences as statistically insignificant, the choice of an appropriate sample size is critical in the design of a clinical program to demonstrate safety and efficacy. When cost is a major factor in implementing a clinical program, the number of patients to be included in the studies may be limited by lack of funds. With fewer patients, a study will be less sensitive. Decreased sensitivity means that the comparative treatments will be relatively more difficult to distinguish statistically if they are, in fact, different.

The problem of choosing a "correct" sample size is related to experimental objectives and the risk (or probability) of coming to an incorrect decision when the experiment and analysis are completed. For simple comparative experiments, certain prior information is required in

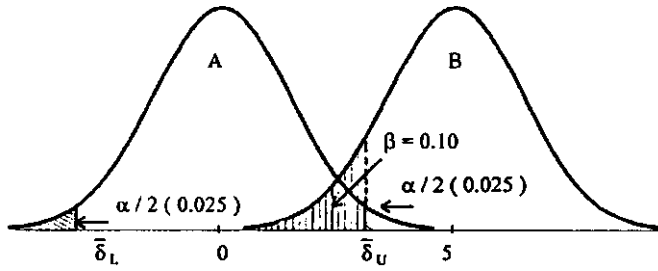
order to compute a sample size that will satisfy the experimental objectives. The following considerations are essential when estimating sample size.

1. The α level must be specified that, in part, determines the difference needed to represent a statistically significant result. To review, the α level is defined as the risk of concluding that treatments differ when, in fact, they are the same. The level of significance is usually (but not always) set at the traditional value of 5%.
2. The β error must be specified for some specified treatment difference, Δ. Beta, β, is the risk (probability) of erroneously concluding that the treatments are not significantly different when, in fact, a difference of size Δ or greater exists. The assessment of β and Δ, the "practically significant" difference, *prior* to the initiation of the experiment, is not easy. Nevertheless, an educated guess is required. β is often chosen to be between 5% and 20%. Hence, one may be willing to accept a 20% (1 in 5) chance of not arriving at a statistically significant difference when the treatments are truly different by an amount equal to (or greater than) Δ. The consequences of committing a β error should be considered carefully. If a true difference of practical significance is missed and the consequence is costly, β should be made very small, perhaps as small as 1%. Costly consequences of missing an effective treatment should be evaluated not only in monetary terms, but should also include public health issues, such as the possible loss of an effective treatment in a serious disease.
3. The difference to be detected, Δ (that difference considered to have practical significance), should be specified as described in (2) above. This difference should not be arbitrarily or capriciously determined, but should be considered carefully with respect to meaningfulness from both a scientific and commercial marketing standpoint. For example, when comparing two formulas for time to 90% dissolution, a difference of one or two minutes might be considered meaningless. A difference of 10 or 20 minutes, however, may have practical consequences in terms of in vivo absorption characteristics.
4. A knowledge of the standard deviation (or an estimate) for the significance test is necessary. If no information on variability is available, an educated guess, or results of studies reported in the literature using related compounds, may be sufficient to give an estimate of the relevant variability. The assistance of a statistician is recommended when estimating the standard deviation for purposes of determining sample size.

To compute the sample size in a comparative experiment, (a) α, (b) β, (c) Δ, and (d) σ must be specified. The computations to determine sample size are described below (Fig. 6.1).

6.2 DETERMINATION OF SAMPLE SIZE FOR SIMPLE COMPARATIVE EXPERIMENTS FOR NORMALLY DISTRIBUTED VARIABLES

The calculation of sample size will be described with the aid of Figure 6.1. This explanation is based on normal distribution or *t* tests. The derivation of sample-size determination may appear complex. The reader not requiring a "proof" can proceed directly to the appropriate formulas below.



$\bar{\delta}_L = -13.7/\sqrt{N}$ $\bar{\delta}_U = -13.7/\sqrt{N} = 5 - 1.28(7)/\sqrt{N}$

Figure 6.1 Scheme to demonstrate calculation of sample size based on α, β, Δ, and σ: $\alpha = 0.05$, $\beta = 0.10$, $\Delta = 5$, $\sigma = 7$; H_0: $\Delta = 0$, H_a: $\Delta = 5$.

6.2.1 Paired-Sample and Single-Sample Tests

We will first consider the case of a paired-sample test where the null hypothesis is that the two treatment means are equal: H_0: $\Delta = 0$. In the case of an experiment comparing a new antihypertensive drug candidate and a placebo, an average difference of 5 mm Hg in blood pressure reduction might be considered of sufficient magnitude to be interpreted as a difference of "practical significance" ($\Delta = 5$). The standard deviation for the comparison was known, equal to 7, based on a large amount of experience with this drug.

In Figure 6.1, the normal curve labeled A represents the distribution of differences with mean equal to 0 and σ equal to 7. This is the distribution under the null hypothesis (i.e., drug and placebo are identical). Curve B is the distribution of differences when the alternative, H_a: $\Delta = 5$,* is true (i.e., the difference between drug and placebo is equal to 5). Note that curve B is identical to curve A except that B is displaced 5 mm Hg to the right. Both curves have the same standard deviation, 7.

With the standard deviation, 7, known, the statistical test is performed at the 5% level as follows [Eq. (5.4)]:

$$Z = \frac{\bar{\delta} - \Delta}{\sigma/\sqrt{N}} = \frac{\bar{\delta} - 0}{7/\sqrt{N}}. \tag{6.1}$$

For a two-tailed test, if the absolute value of Z is 1.96 or greater, the difference is significant. According to Eq. (6.1), to obtain the significance

$$|\bar{\delta}| \geq \frac{\sigma Z}{\sqrt{N}} = \frac{7(1.96)}{\sqrt{N}} = \frac{13.7}{\sqrt{N}}. \tag{6.2}$$

Therefore, values of $\bar{\delta}$ equal to or greater than $13.7/\sqrt{N}$ (or equal to or less than $-13.7/\sqrt{N}$) will lead to a declaration of significance. These points are designated as $\bar{\delta}_L$ and $\bar{\delta}_U$ in Figure 6.1, and represent the cutoff points for statistical significance at the 5% level; that is, observed differences equal to or more remote from the mean than these values result in "statistically significant differences."

If curve B is the true distribution (i.e., $\Delta = 5$), an observed mean difference greater than $13.7/\sqrt{N}$ (or less than $-13.7/\sqrt{N}$) will result in the correct decision; H_0 will be rejected and we conclude that a difference exists. If $\Delta = 5$, observations of a mean difference between $13.7/\sqrt{N}$ and $-13.7/\sqrt{N}$ will lead to an *incorrect decision*, the acceptance of H_0 (no difference) (Fig. 6.1). By definition, the probability of making this *incorrect* decision is equal to β.

In the present example, β will be set at 10%. In Figure 6.1, β is represented by the area in curve B below $13.7/\sqrt{N}(\delta_U)$, equal to 0.10. (This area, β, represents the probability of accepting H_0 if $\Delta = 5$.)

We will now compute the value of $\bar{\delta}$ that cuts off 10% of the area in the lower tail of the normal curve with a mean of 5 and a standard deviation of 7 (curve B in Figure 6.1). Table IV.2 shows that 10% of the area in the standard normal curve is below -1.28. The value of $\bar{\delta}$ (mean difference in blood pressure between the two groups) that corresponds to a given value of Z (-1.28, in this example) is obtained from the formula for the Z transformation [Eq. (3.14)] as follows:

$$\bar{\delta} = \Delta + Z_\beta \left(\frac{\sigma}{\sqrt{N}}\right)$$

$$Z_\beta = \frac{\bar{\delta} - \Delta}{\sigma/\sqrt{N}}. \tag{6.3}$$

Applying Eq. (6.3) to our present example, $\bar{\delta} = 5 - 1.28(7/\sqrt{N})$. The value of $\bar{\delta}$ in Eqs. (6.2) and (6.3) is identically the same, equal to $\bar{\delta}_U$. This is illustrated in Figure 6.1.

* Δ is considered to be the *true* mean difference, similar to μ. $\bar{\delta}$ will be used to denote the *observed* mean difference.

Table 6.1 Sample Size as a Function of Beta with $\Delta = 5$ and $\sigma = 7$: Paired Test ($\alpha = 0.05$)

Beta (%)	Sample size, *N*
1	36
5	26
10	21
20	16

From Eq. (6.2), $\bar{\delta}_U = 13.7/\sqrt{N}$, satisfying the definition of α. From Eq. (6.3), $\bar{\delta}_U = 5 - 1.28(7)/\sqrt{N}$, satisfying the definition of β. We have two equations in two unknowns ($\bar{\delta}_U$ and N), and N is evaluated as follows:

$$\frac{13.7}{\sqrt{N}} = 5 - \frac{1.28(7)}{\sqrt{N}}$$

$$N = \frac{(13.7 + 8.96)^2}{5^2} = 20.5 \cong 21.$$

In general, Eqs. (6.2) and (6.3) can be solved for N to yield the following equation:

$$N = \left(\frac{\sigma}{\Delta}\right)^2 (Z_\alpha + Z_\beta)^2, \tag{6.4}$$

where Z_α and Z_β[†] are the appropriate normal deviates obtained from Table IV.2. In our example, $N = (7/5)^2(1.96 + 1.28)^2 \cong 21$. A sample size of 21 will result in a statistical test with 90% power ($\beta = 10\%$) against an alternative of 5, at the 5% level of significance. Table 6.1 shows how the choice of β can affect the sample size for a test at the 5% level with $\Delta = 5$ and $\sigma = 7$.

The formula for computing the sample size if the standard deviation is known [Eq. (6.4)] is appropriate for a paired-sample test or for the test of a mean from a *single population.* For example, consider a test to compare the mean drug content of a sample of tablets to the labeled amount, 100 mg. The two-sided test is to be performed at the 5% level. Beta is designated as 10% for a difference of -5 mg (95 mg potency or less). That is, we wish to have a power of 90% to detect a difference from 100 mg if the true potency is 95 mg or less. If σ is equal to 3, how many tablets should be assayed? Applying Eq. (6.4), we have

$$N = \left(\frac{3}{5}\right)^2 (1.96 + 1.28)^2 = 3.8.$$

Assaying four tablets will satisfy the α and β probabilities. Note that $Z = 1.28$ cuts off 90% of the area under curve B (the "alternative" curve) in Figure 6.2, leaving 10% (β) of the area in the upper tail of the curve. Table 6.2 shows values of Z_α and Z_β for various levels of α and β to be used in Eq. (6.4). In this example, and most examples in practice, β is based on one tail of the normal curve. The other tail contains an insignificant area relating to β (the right side of the normal curve, B, in Fig. 6.1)

Equation (6.4) is correct for computing the sample size for a paired- or one-sample test if the standard deviation is known.

In most situations, the standard deviation is unknown and a prior estimate of the standard deviation is necessary in order to calculate sample size requirements. In this case, the estimate of the standard deviation replaces σ in Eq. (6.4), but the calculation results in an answer that is slightly too small. The underestimation occurs because the values of Z_α and Z_β are smaller than

[†] Z_β is taken as the positive value of Z in this formula.

Table 6.2 Values of Z_α and Z_β for Sample-Size Calculations

	Z_α		
	One sided	**Two sided**	Z_β^a
1%	2.32	2.58	2.32
5%	1.65	1.96	1.65
10%	1.28	1.65	1.28
20%	0.84	1.28	0.84

[a]The value of β is for a single specified alternative. For a two-sided test, the probability of rejection of the alternative, if true, (accept H_a) is virtually all contained in the tail nearest the alternative mean.

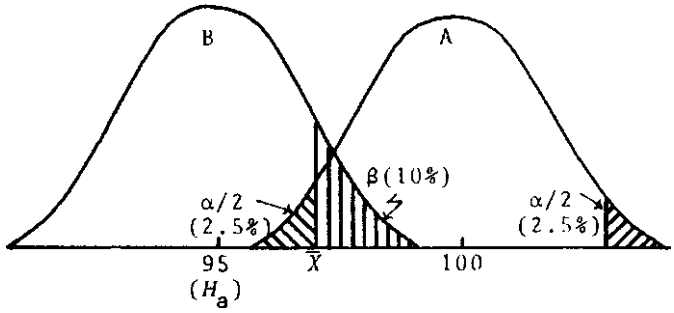


Figure 6.2 Illustration of the calculation of *N* for tablet assays. $\overline{X} = 95 + \sigma Z_\beta/\sqrt{N} = 100 - \sigma Z_\alpha/\sqrt{N}$.

the corresponding *t* values that should be used in the formula when the standard deviation is unknown. The situation is somewhat complicated by the fact that the value of *t* depends on the sample size (d.f.), which is yet unknown. The problem can be solved by an iterative method, but for practical purposes, one can use the appropriate values of *Z* to compute the sample size [as in Eq. (6.4)] and add on a few extra samples (patients, tablets, etc.) to compensate for the use of *Z* rather than *t*. Guenther has shown that the simple addition of $0.5Z_\alpha^2$, which is equal to approximately 2 for a two-sided test at the 5% level, results in a very close approximation to the correct answer [3]. In the problem illustrated above (tablet assays), if the standard deviation were *unknown* but *estimated* as being equal to 3 based on previous experience, a better estimate of the sample size would be $N + 0.5Z_\alpha^2 = 3.8 + 0.5(1.96)^2 \cong 6$ tablets.

6.2.2 Determination of Sample Size for Comparison of Means in Two Groups

For a two independent groups test (parallel design), with the standard deviation known and equal number of observations per group, the formula for *N* (where *N* is the sample size for each group) is

$$N = 2\left(\frac{\sigma}{\Delta}\right)^2 (Z_\alpha + Z_\beta)^2. \tag{6.5}$$

If the standard deviation is unknown and a prior estimate is available (s.d.), substitute s.d. for σ in Eq. (6.5) and compute the sample size; but add on $0.25Z_\alpha^2$ to the sample size for each group.

Example 1: This example illustrates the determination of the sample size for a two independent groups (two-sided test) design. Two variations of a tablet formulation are to be compared with regard to dissolution time. All ingredients except for the lubricating agent were the same in these two formulations. In this case, a decision was made that if the formulations differed by 10 minutes or more to 80% dissolution, it would be extremely important that the experiment shows a statistically significant difference between the formulations. Therefore, the pharmaceutical scientist decided to fix the β error at 1% in a statistical test at the traditional 5% level. Data were available from dissolution tests run during the development of formulations of the drug

and the standard deviation was *estimated* as 5 minutes. With the information presented above, the sample size can be determined from Eq. (6.5). We will add on $0.25Z_{\alpha}^2$ samples to the answer because the standard deviation is unknown.

$$N = 2\left(\frac{5}{10}\right)^2 (1.96 + 2.32)^2 + 0.25(1.96)^2 = 10.1.$$

The study was performed using 12 tablets from each formulation rather than the 10 or 11 suggested by the answer in the calculation above. Twelve tablets were used because the dissolution apparatus could accommodate six tablets per run.

Example 2: A bioequivalence study was being planned to compare the bioavailability of a final production batch to a previously manufactured pilot-sized batch of tablets that were made for clinical studies. Two parameters resulting from the blood-level data would be compared: area under the plasma level versus time curves (AUC) and peak plasma concentration (C_{max}). The study was to have 80% power ($\beta = 0.20$) to detect a difference of 20% or more between the formulations. The test is done at the usual 5% level of significance. Estimates of the standard deviations of the *ratios* of the values of each of the parameters [(final product)/(pilot batch)] were determined from a small pilot study. The standard deviations were different for the parameters. Since the researchers could not agree that one of the parameters was clearly critical in the comparison, they decided to use a "maximum" number of patients based on the variable with the largest relative variability. In this example, C_{max} was most variable, the ratio having a standard deviation of approximately 0.30. Since the design and analysis of the bioequivalence study is a variation of the paired t test, Eq. (6.4) was used to calculate the sample size, adding on $0.5Z_{\alpha}^2$, as recommended previously.

$$\begin{aligned} N &= \left(\frac{\sigma}{\Delta}\right)^2 (Z_{\alpha} + Z_{\beta})^2 + 0.5(Z_{\alpha}^2) \\ &= \left(\frac{0.3}{0.2}\right)^2 (1.96 + 0.84)^2 + 0.5(1.96)^2 = 19.6. \end{aligned} \tag{6.6}$$

Twenty subjects were used for the comparison of the bioavailabilities of the two formulations.

For sample-size determination for bioequivalence studies using FDA recommended designs, see Table 6.5 and section 11.4.4.

Sometimes the sample sizes computed to satisfy the desired α and β errors can be inordinately large when time and cost factors are taken into consideration. Under these circumstances, a compromise must be made—most easily accomplished by relaxing the α and β requirements[‡] (Table 6.1). The consequence of this compromise is that probabilities of making an incorrect decision based on the statistical test will be increased. Other ways of reducing the required sample size are (a) to increase the precision of the test by improving the assay methodology or carefully controlling extraneous conditions during the experiment, for example, or (b) to compromise by increasing Δ, that is, accepting a larger difference that one considers to be of practical importance.

Table 6.3 gives the sample size for some representative values of the ratio σ/Δ, α, and β, where the s.d. (s) is estimated.

6.3 DETERMINATION OF SAMPLE SIZE FOR BINOMIAL TESTS

The formulas for calculating the sample size for comparative binomial tests are similar to those described for normal curve or t tests. The major difference is that the value of σ^2, which is assumed to be the same under H_0 and H_a in the two-sample independent groups t or Z tests, is different for the distributions under H_0 and H_a in the binomial case. This difference occurs because σ^2 is dependent on P, the probability of success, in the binomial. The value of P will

[‡] In practice, α is often fixed by regulatory considerations and β is determined as a compromise.

Table 6.3 Sample Size Needed for Two-Sided *t* Test with Standard Deviation Estimated

	One-sample test								Two-sample test with *N* units per group							
	Alpha = 0.05				Alpha = 0.01				Alpha = 0.05				Alpha = 0.01			
	Beta =				Beta =				Beta =				Beta =			
Estimated *S*/Δ	**0.01**	**0.05**	**0.10**	**0.20**	**0.01**	**0.05**	**0.10**	**0.20**	**0.01**	**0.05**	**0.10**	**0.20**	**0.01**	**0.05**	**0.10**	**0.20**
4.0	296	211	170	128	388	289	242	191	588	417	337	252	770	572	478	376
2.0	76	54	44	34	100	75	63	51	148	106	86	64	194	145	121	96
1.5	44	32	26	20	58	54	37	30	84	60	49	37	110	82	69	55
1.0	21	16	13	10	28	22	19	16	38	27	23	17	50	38	32	26
0.8	14	11	9	8	19	15	13	11	25	18	15	12	33	25	21	17
0.67	11	8	7	6	15	12	11	9	18	13	11	9	24	18	15	13
0.5	7	6	5	4	10	8	8	7	11	8	7	6	14	11	10	8
0.4	6	5	4	4	8	7	6	6	8	6	5	4	10	8	7	6
0.33	5	4	4	3	7	6	6	5	6	5	4	4	8	6	6	5

be different depending on whether H_0 or H_a represents the true situation. The appropriate formulas for determining sample size for the one- and two-sample tests are

One-sample test

$$N = \frac{1}{2}\left[\frac{p_0q_0 + p_1q_1}{\Delta^2}\right](Z_\alpha + Z_\beta)^2, \tag{6.7}$$

where $\Delta = p_1 - p_0$; p_1 is the proportion that would result in a meaningful difference, and p_0 is the hypothetical proportion under the null hypothesis.

Two-sample test

$$N = \left[\frac{p_1q_1 + p_2q_2}{\Delta^2}\right](Z_\alpha + Z_\beta)^2, \tag{6.8}$$

where $\Delta = p_1 - p_2$; p_1 and p_2 are prior estimates of the proportions in the experimental groups. The values of Z_α and Z_β are the same as those used in the formulas for the normal curve or t tests. N is the sample size for each group. If it is not possible to estimate p_1 and p_2 prior to the experiment, one can make an educated guess of a meaningful value of Δ and set p_1 and p_2 both equal to 0.5 in the *numerator* of Eq. (6.8). This will maximize the sample size, resulting in a conservative estimate of sample size.

Fleiss [4] gives a fine discussion of an approach to estimating Δ, the practically significant difference, when computing the sample size. For example, one approach is first to estimate the proportion for the more well-studied treatment group. In the case of a comparative clinical study, this could very well be a standard treatment. Suppose this treatment has shown a success rate of 50%. One might argue that if the comparative treatment is additionally successful for 30% of the patients who do not respond to the standard treatment, then the experimental treatment would be valuable. Therefore, the success rate for the experimental treatment should be 50% + 0.3 (50%) = 65% to show a practically significant difference. Thus, p_1 would be equal to 0.5 and p_2 would be equal to 0.65.

Example 3: A reconciliation of quality control data over several years showed that the proportion of unacceptable capsules for a stable encapsulation process was 0.8% (p_0). A sample size for inspection is to be determined so that if the true proportion of unacceptable capsules is equal to or greater than 1.2% ($\Delta = 0.4\%$), the probability of detecting this change is 80% ($\beta = 0.2$). The comparison is to be made at the 5% level using a *one-sided* test. According to Eq. (6.7),

$$\begin{aligned} N &= \frac{1}{2}\left[\frac{0.008 \cdot 0.992 + 0.012 \cdot 0.988}{(0.008 - 0.012)^2}\right](1.65 + 0.84)^2 \\ &= \frac{7670}{2} \\ &= 3835. \end{aligned}$$

The large sample size resulting from this calculation is typical of that resulting from binomial data. If 3835 capsules are too many to inspect, α, β, and/or Δ must be increased. In the example above, management decided to increase α. This is a conservative decision in that more good batches would be "rejected" if α is increased; that is, the increase in α results in an increased probability of rejecting good batches, those with 0.8% unacceptable or less.

Example 4: Two antibiotics, a new product and a standard product, are to be compared with respect to the two-week cure rate of a urinary tract infection, where a cure is bacteriological evidence that the organism no longer appears in urine. From previous experience, the cure rate for the standard product is estimated at 80%. From a practical point of view, if the new product shows an 85% or better cure rate, the new product can be considered superior. The marketing

division of the pharmaceutical company felt that this difference would support claims of better efficacy for the new product. This is an important claim. Therefore, β is chosen to be 1% (power = 99%). A two-sided test will be performed at the 5% level to satisfy FDA guidelines. The test is two-sided because, a priori, the new product is not known to be better or worse than the standard. The calculation of sample size to satisfy the conditions above makes use of Eq. (6.8); here $p_1 = 0.8$ and $p_2 = 0.85$.

$$N = \left[\frac{0.08 \cdot 0.2 + 0.85 \cdot 0.15}{(0.80 - 0.85)^2}\right](1.96 + 2.32)^2 = 2107.$$

The trial would have to include 4214 patients, 2107 on each drug, to satisfy the α and β risks of 0.05 and 0.01, respectively. If this number of patients is greater than that can be accommodated, the β error can be increased to 5% or 10%, for example. A sample size of 1499 per group is obtained for a β of 5%, and 1207 patients per group for β equal to 10%.

Although Eq. (6.8) is adequate for computing the sample size for most situations, the calculation of N can be improved by considering the continuity correction [4]. This would be particularly important for small sample sizes

$$N' = \left[\frac{N}{4}\right]\left[1 + \sqrt{1 + \frac{8}{(N\,|p_2 - p_1|)}}\right]^2,$$

where N is the sample size computed from Eq. (6.8) and N' is the corrected sample size. In the example, for $\alpha = 0.05$ and $\beta = 0.01$, the corrected sample size is

$$N' = \left[\frac{2107}{4}\right]\left[1 + \sqrt{1 + \frac{8}{(2107\,|0.80 - 0.85|)}}\right]^2 = 2186.$$

6.4 DETERMINATION OF SAMPLE SIZE TO OBTAIN A CONFIDENCE INTERVAL OF SPECIFIED WIDTH

The problem of estimating the number of samples needed to estimate the mean with a known precision by means of the confidence interval is easily solved by using the formula for the confidence interval (see sect. 5.1). This approach has been used as an aid in predicting election results based on preliminary polls where the samples are chosen by simple random sampling. For example, one may wish to estimate the proportion of voters who will vote for candidate A within 1% of the actual proportion.

We will consider the application of this problem to the estimation of proportions. In quality control, one can closely estimate the true proportion of percent defects to any given degree of precision. In a clinical study, a suitable sample size may be chosen to estimate the true proportion of successes within certain specified limits. According to Eq. (5.3), a two-sided confidence interval with confidence coefficient p for a proportion is

$$\hat{p} \pm Z\sqrt{\frac{\hat{p}\hat{q}}{N}}. \tag{6.3}$$

To obtain a 99% confidence interval with a width of 0.01 (i.e., construct an interval that is within ±0.005 of the observed proportion, $\hat{p} \pm 0.005$),

$$Z_p\sqrt{\frac{\hat{p}\hat{q}}{N}} = 0.005$$

or

$$N = \frac{Z_p^2(\hat{p}\hat{q})}{(W/2)^2} \tag{6.9}$$

$$N = \frac{(2.58)^2(\hat{p}\hat{q})}{(0.005)^2}.$$

A more exact formula for the sample size for small values of N is given in Ref. [5].

Example 5: A quality control supervisor wishes to have an estimate of the proportion of tablets in a batch that weigh between 195 and 205 mg, where the proportion of tablets in this interval is to be estimated within ± 0.05 ($W = 0.10$). How many tablets should be weighed? Use a 95% confidence interval.

To compute N, we must have an estimate of $\hat{p}$ [see Eq. (6.9)]. If $\hat{p}$ and $\hat{q}$ are chosen to be equal to 0.5, N will be at a maximum. Thus, if one has no inkling as to the magnitude of the outcome, using $\hat{p} = 0.5$ in Eq. (6.9) will result in a sufficiently large sample size (probably, too large). Otherwise, estimate $\hat{p}$ and $\hat{q}$ based on previous experience and knowledge. In the present example from previous experience, approximately 80% of the tablets are expected to weigh between 195 and 205 mg ($\hat{p} = 0.8$). Applying Eq. (6.9),

$$N = \frac{(1.96)^2(0.8)(0.2)}{(0.10/2)^2} = 245.9.$$

A total of 246 tablets should be weighed. In the actual experiment, 250 tablets were weighed, and 195 of the tablets (78%) weighed between 195 and 205 mg. The 95% confidence interval for the true proportion, according to Eq. (5.3), is

$$p \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{N}} = 0.78 \pm 1.96\sqrt{\frac{(0.78)(0.22)}{250}} = 0.78 \pm 0.051.$$

The interval is slightly greater than $\pm 5\%$ because p is somewhat less than 0.8 (pq is larger for $p = 0.78$ than for $p = 0.8$). Although 5.1% is acceptable, to ensure a sufficient sample size, in general, one should estimate p closer to 0.5 in order to cover possible poor estimates of p.

If $\hat{p}$ had been chosen equal to 0.5, we would have calculated

$$N = \frac{(1.96)^2(0.5)(0.5)}{(0.10/2)^2} = 384.2.$$

Example 6: A new vaccine is to undergo a nationwide clinical trial. An estimate is desired of the proportion of the population that would be afflicted with the disease after vaccination. A good guess of the expected proportion of the population diseased without vaccination is 0.003. Pilot studies show that the incidence will be about 0.001 (0.1%) after vaccination. What size sample is needed so that the width of a 99% confidence interval for the proportion diseased in the vaccinated population should be no greater than 0.0002? To ensure that the sample size is sufficiently large, the value of p to be used in Eq. (6.9) is chosen to be 0.0012, rather than the expected 0.0010.

$$N = \frac{(2.58)^2(0.9988)(0.0012)}{(0.0002/2)^2} = 797{,}809.$$

The trial will have to include approximately 800,000 subjects in order to yield the desired precision.

6.5 POWER

Power is the probability that the statistical test results in rejection of H_0 when a specified alternative is true. The "stronger" the power, the better the chance that the null hypothesis will be rejected (i.e., the test results in a declaration of "significance") when, in fact, H_0 is false. The larger the power, the more sensitive is the test. *Power* is defined as $1 - \beta$. The larger the β error, the weaker is the power. Remember that β is an error resulting from *accepting* H_0 *when* H_0 *is false.* Therefore, $1 - \beta$ is the probability of *rejecting* H_0 *when* H_0 *is false.*

From an idealistic point of view, the power of a test should be calculated *before* an experiment is conducted. In addition to defining the properties of the test, power is used to help compute the sample size, as discussed above. Unfortunately, many experiments proceed without consideration of power (or β). This results from the difficulty of choosing an appropriate value of β. There is no traditional value of β to use, as is the case for α, where 5% is usually used. Thus, the power of the test is often computed after the experiment has been completed.

Power is best described by diagrams such as those shown previously in this chapter (Figs. 6.1 and 6.2). In these figures, β is the area of the curves represented by the alternative hypothesis that is included in the region of acceptance defined by the null hypothesis.

The concept of power is also illustrated in Figure 6.3. To illustrate the calculation of power, we will use data presented for the test of a new antihypertensive agent (sect. 6.2), a paired sample test, with $\sigma = 7$ and $H_0 : \Delta = 0$. The test is performed at the 5% level of significance. Let us suppose that the sample size is limited by cost. The sponsor of the test had sufficient funds to pay for a study that included only *12 subjects.* The design described earlier in this chapter (sect. 6.2) used 26 patients with β specified equal to 0.05 (power = 0.95). With 12 subjects, the power will be considerably less than 0.95. The following discussion shows how power is calculated.

The cutoff points for statistical significance (which specify the critical region) are defined by α, N, and σ. Thus, the values of $\bar{\delta}$ that will lead to a significant result for a two-sided test are as follows:

$$Z = \frac{|\bar{\delta}|}{\sigma/\sqrt{N}}$$

$$\bar{\delta} = \frac{\pm Z\sigma}{\sqrt{N}}.$$

In our example, $Z = 1.96$ ($\alpha = 0.05$), $\sigma = 7$, and $N = 12$.

$$\bar{\delta} = \frac{\pm(1.96)(7)}{\sqrt{12}} = \pm 3.96.$$

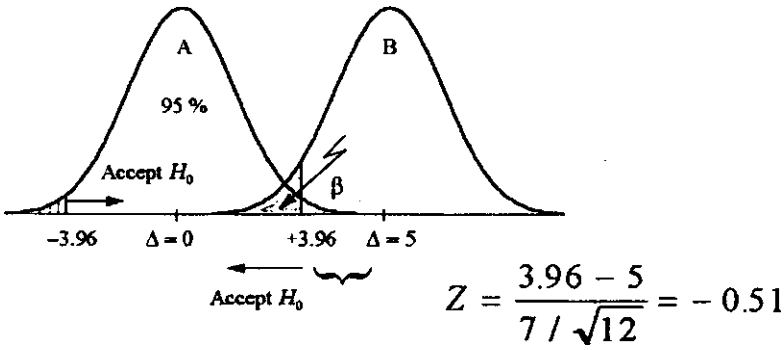


Figure 6.3 Illustration of beta or power $(1 - \beta)$.

Values of $\bar{\delta}$ greater than 3.96 or less than -3.96 will lead to the decision that the products differ at the 5% level. Having defined the values of $\bar{\delta}$ that will lead to rejection of H_0, we obtain the power for the alternative, H_a: $\Delta = 5$, by computing the probability that an average result, $\bar{\delta}$, will be greater than 3.96, if H_a is true (i.e., $\Delta = 5$).

This concept is illustrated in Figure 6.3. Curve B is the distribution with mean equal to 5 and $\sigma = 7$. If curve B is the true distribution, the probability of observing a value of $\bar{\delta}$ below 3.96 is the probability of accepting H_0 if the alternative hypothesis is true ($\Delta = 5$). This is the definition of β. This probability can be calculated using the Z transformation.

$$Z = \frac{3.96 - 5}{7/\sqrt{12}} = -0.51.$$

Referring to Table IV.2, the area below $+3.96$ ($Z = -0.51$) for curve B is approximately 0.31. The power is $1 - \beta = 1 - 0.31 = 0.69$. The use of 12 subjects results in a power of 0.69 to "detect" a difference of $+5$ compared to the 0.95 power to detect such a difference when 26 subjects were used. A power of 0.69 means that if the *true difference were 5 mm Hg*, the statistical test will result in *significance with a probability of 69%*; 31% of the time, such a test will result in acceptance of H_0.

A *power curve* is a plot of the power, $1 - \beta$, versus alternative values of Δ. Power curves can be constructed by computing β for several alternatives and drawing a smooth curve through these points. For a two-sided test, the power curve is symmetrical around the hypothetical mean, $\Delta = 0$, in our example. The power is equal to α when the alternative is equal to the hypothetical mean under H_0. Thus, the power is 0.05 when $\Delta = H_0$ (Fig. 6.4) in the power curve. The power curve for the present example is shown in Figure 6.4.

The following conclusions may be drawn concerning the *power* of a test if α is kept constant:

1. The larger the sample size, the larger the power.
2. The larger the difference to be detected (H_a), the larger the power. A large sample size will be needed in order to have strong power to detect a small difference.
3. The larger the variability (s.d.), the weaker the power.
4. If α is increased, power is increased (β is decreased) (Fig. 6.3). An increase in α (e.g., 10%) results in a smaller Z. The cutoff points are shorter, and the area of curve B below the cutoff point is smaller.

Power is a function of N, Δ, σ, and α.

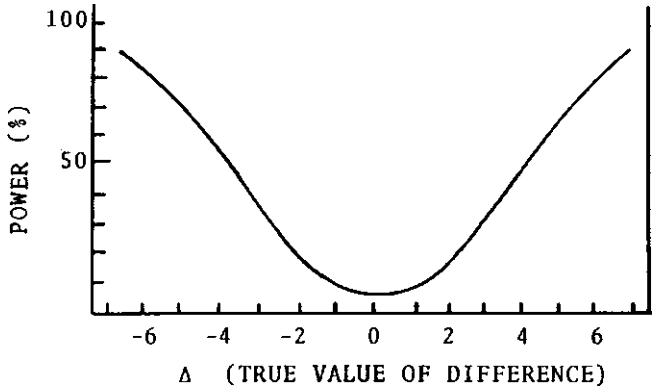


Figure 6.4 Power curve for $N = 12$, $\alpha = 0.05$, $\sigma = 7$, and H_0: $\Delta = 0$.

A simple way to compute the approximate power of a test is to use the formula for sample size [Eqs. (6.4) and (6.5). for example] and solve for Z_β. In the previous example, a single sample or a paired test, Eq. (6.4) is appropriate:

$$N = \left(\frac{\sigma}{\Delta}\right)^2 (Z_\alpha + Z_\beta)^2 \tag{6.4}$$

$$Z_\beta = \frac{\Delta}{\sigma}\sqrt{N} - Z_\alpha. \tag{6.10}$$

Once having calculated Z_β, the probability determined directly from Table IV.2 is equal to the power, $1 - \beta$. See the discussion and examples below.

In the problem discussed above, applying Eq. (6.10) with $\Delta = 5$, $\sigma = 7$, $N = 12$, and $Z_\alpha = 1.96$,

$$Z_\beta = \frac{5}{7}\sqrt{12} - 1.96 = 0.51.$$

According to the notation used for Z (Table 6.2), β is the area above Z_β. Power is the area below Z_β(power $= 1 - \beta$). In Table IV.2, the area above $Z = 0.51$ is approximately 31%. The power is $1 - \beta$. Therefore, the power is 69%.§

If N is small and the variance is unknown, appropriate values of t should be used in place of Z_α and Z_β. Alternatively, we can adjust N by subtracting $0.5Z_\alpha^2$ or $0.25Z_\alpha^2$ from the actual sample size for a one- or two-sample test, respectively. The following examples should make the calculations clearer.

Example 7: A bioavailability study has been completed in which the ratio of the AUCs for two comparative drugs was submitted as evidence of bioequivalence. The FDA asked for the power of the test as part of their review of the submission. (Note that this analysis is different from that presently required by FDA.) The null hypothesis for the comparison is H_0: $R = 1$, where R is the true average ratio. The test was two-sided with α equal to 5%. Eighteen subjects took each of the two comparative drugs in a paired-sample design. The standard deviation was calculated from the final results of the study, and was equal to 0.3. The power is to be determined for a difference of 20% for the comparison. This means that if the test product is truly more than 20% greater or smaller than the reference product, we wish to calculate the probability that the ratio will be judged to be significantly different from 1.0. The value of Δ to be used in Eq. (6.10) is 0.2.

$$Z_\beta = \frac{0.2\sqrt{16}}{0.3} - 1.96 = 0.707.$$

Note that the value of N is taken as 16. This is the inverse of the procedure for determining sample size, where $0.5Z_\alpha^2$ was added to N. Here we subtract $0.5Z_\alpha^2$ (approximately 2) from N; $18 - 2 = 16$. According to Table IV.2, the area corresponding to $Z = 0.707$ is approximately 0.76. Therefore, the power of this test is 76%. That is, if the true difference between the formulations is 20%, a significant difference will be found between the formulations 76% of the time. This is very close to the 80% power that was recommended before current FDA guidelines were implemented for bioavailability tests (where $\Delta = 0.2$).

Example 8: A drug product is prepared by two different methods. The average tablet weights of the two batches are to be compared, weighing 20 tablets from each batch. The average weights of the two 20-tablet samples were 507 and 511 mg. The pooled standard deviation was calculated to be 12 mg. The director of quality control wishes to be "sure" that if the average weights truly differ by 10 mg or more, the statistical test will show a significant difference, when

§ The value corresponding to Z in Table IV.2 gives the power directly. In this example, the area in the table corresponding to a Z of 0.51 is approximately 0.69.

he was asked, "How sure?", he said 95% sure. This can be translated into a β of 5% or a power of 95%. This is a *two independent groups* test. Solving for Z_β from Eq. (6.5), we have

$$Z_\beta = \frac{\Delta}{\sigma}\sqrt{\frac{N}{2}} - Z_\alpha$$
$$= \frac{10}{12}\sqrt{\frac{19}{2}} - 1.96 = 0.609. \tag{6.11}$$

As discussed above, the value of N is taken as 19 rather than 20, by subtracting $0.25Z_\alpha^2$ from N for the two-sample case. Referring to Table IV.2, we note that the power is approximately 73%. The experiment does not have sufficient power according to the director's standards. To obtain the desired power, we can increase the sample size (i.e., weigh more tablets). (See Exercise Problem 10.)

6.6 SAMPLE SIZE AND POWER FOR MORE THAN TWO TREATMENTS (ALSO SEE CHAP. 8)

The problem of computing power or sample size for an experiment with more than two treatments is somewhat more complicated than the relatively simple case of designs with two treatments. The power will depend on the number of treatments and the form of the null and alternative hypotheses. Dixon and Massey [5] present a simple approach to determining power and sample size. The following notation will be used in presenting the solution to this problem.

Let $M_1, M_2, M_3 \ldots M_k$ be the hypothetical population means of the k treatments. The null hypothesis is $M_1 = M_2 = \mathrm{M}_3 = M_k$. As for the two sample cases, we must specify the alternative values of M_i. The alternative means are expressed as a grand mean, $M_t \pm$ some deviation, D_i, where $\sum(D_i) = 0$. For example, if three treatments are compared for pain, Active A, Active B, and Placebo (P), the values for the alternative hypothesized means, based on a VAS scale for pain relief, could be $75 + 10$ (85), $75 + 10$ (85), and $75 - 20$ (55) for the two actives and placebo, respectively. The sum of the deviations from the grand mean, 75, is $10 + 10 - 20 = 0$. The power is computed based on the following equation:

$$\psi^2 = \frac{\sum (M_i - M_t)^2/k}{S^2/n}, \tag{6.12}$$

where n is the number of observations in each treatment group (n is the same for each treatment) and S^2 is the common variance. The value of ψ^2 is referred to Table 6.4 to estimate the required sample size.

Consider the following example of three treatments in a study measuring the analgesic properties of two actives and a placebo as described above. Fifteen subjects are in each treatment group and the variance is 1000. According to Eq. (6.12),

$$\psi^2 = \frac{\left\{(85-75)^2 + (85-75)^2 + (55-75)^2\right\}/3}{1000/15} = 3.0.$$

Table 6.4 gives the approximate power for various values of ψ, at the 5% level, as a function of the number of treatment groups and the d.f. for error for 3 and 4 treatments. (More detailed tables, in addition to graphs, are given in Dixon and Massey [5].) Here, we have 42 d.f. and three treatments with $\psi = \sqrt{3} = 1.73$. The power is approximately 0.72 by simple linear interpolation (42 d.f. for $\psi = 1.7$). The correct answer with more extensive tables is closer to 0.73.

Table 6.4 Factors for Computing Power for Analysis of Variance

d.f. error	ψ	Power
Alpha = 0.05, $k = 3$		
10	1.6	0.42
	2.0	0.76
	2.4	0.80
	3.0	0.984
20	1.6	0.62
	1.92	0.80
	2.00	0.83
	3.0	>0.99
30	1.6	0.65
	1.9	0.80
	2.0	0.85
	3.0	>0.99
60	1.6	0.67
	1.82	0.80
	2.0	0.86
	3.0	>0.99
inf	1.6	0.70
	1.8	0.80
	2.0	0.88
	3.0	>0.99
alpha = 0.05, $k = 4$		
10	1.4	0.48
	2.0	0.80
	2.6	0.96
20	1.4	0.56
	2.0	0.88
	2.6	986
30	1.4	0.59
	2.0	0.90
	2.6	>0.99
60	1.4	0.61
	2.0	0.92
	2.6	>0.99
inf	1.4	0.65
	2.0	0.94
	2.6	>0.99

Table 6.4 can also be used to determine sample size. For example, how many patients per treatment group are needed to obtain a power of 0.80 in the above example? Applying Eq. (6.12),

$$\frac{\{(85-75)^2+(85-75)^2+(55-75)^2\}/3}{1000/n} = \psi^2.$$

Solve for ψ^2

$$\psi^2 = 0.2n.$$

We can calculate n by trial and error. For example, with $N = 20$,

$$0.2N = 4 = \psi^2 \quad \text{and} \quad \psi = 2.$$

For $\psi = 2$ and $N = 20$ (d.f. = 57), the power is approximately 0.86 (for d.f. = 60, power 0.86). For $N = 15$ (d.f. = 42, $\psi = \sqrt{3}$), we have calculated (above) that the power is approximately 0.72. A sample size of between 15 and 20 patients per treatment group would give a power of 0.80. In this example, we might guess that 17 patients per group would result in approximately 80% power. Indeed, more exact tables show that a sample size of 17($\psi = \sqrt{(0.2 \times 17)} = 1.85$) corresponds to a power of 0.79.

The same approach can be used for two-way designs, using the appropriate error term from the analysis of variance.

6.7 SAMPLE SIZE FOR BIOEQUIVALENCE STUDIES (ALSO SEE CHAP. 11)

In its early evolution, bioequivalence was based on the acceptance or rejection of a hypothesis test. Sample sizes could then be determined by conventional techniques as described in section 6.2. Because of inconsistencies in the decision process based on this approach, the criteria for acceptance was changed to a two-sided 90% confidence interval, or equivalently, two one-sided t test, where the hypotheses are $(\mu_1/\mu_2) < 0.8$ and $(\mu_1/\mu_2) > 1.25$ versus the alternative of $0.8 < (\mu_1/\mu_2) < 1.25$. This test is based on the antilog of the difference between the averages of the log-transformed parameters (the geometric mean). This test is equivalent to a two-sided 90% confidence interval for the ratio of means falling in the interval 0.80 to 1.25 in order to accept the hypothesis of equivalence. Again, for the currently accepted log-transformed data, the 90% confidence interval for the antilog of the difference between means must lie between 0.80 and 1.25, that is, $0.8 < \text{antilog}\,(\mu_1/\mu_2) < 1.25$. The sample-size determination in this case is not as simple as the conventional determination of sample size described earlier in this chapter. The method for sample-size determination for nontransformed data has been published by Phillips [6] along with plots of power as a function of sample size, relative standard deviation (computed from the ANOVA), and treatment differences. Although the theory behind this computation is beyond the scope of this book, Chow and Liu [7] give a simple way of approximating the power and sample size. The sample size for each sequence group is approximately

$$N = (t_{\alpha,\,2N-2} + t_{\beta,\,2N-2})^2 \left[\frac{\text{CV}}{(\text{V} - \delta)}\right]^2, \tag{6.13}$$

where N is the number of subjects per sequence, t the appropriate value from the t distribution, α the significance level (usually 0.10), $1 - \beta$ the power (usually 0.8), CV the coefficient of variation, V the bioequivalence limit, and δ the difference between products.

One would have to have an approximation of the magnitude of the required sample size in order to approximate the t values. For example, suppose that RSD = 0.20, $\delta = 0.10$, power is 0.8, and an initial approximation of the sample size is 20 per sequence (a total of 40 subjects). Applying Eq. (6.13)

$$n = (1.69 + 0.85)^2[0.20/(0.20 - 0.10)]^2 = 25.8.$$

Use a total of 52 subjects. This agrees closely with Phillip's more exact computations. Dilletti et al. [8] have published a method for determining sample size based on the log-transformed variables, which is the currently preferred method. Table 6.5 showing sample sizes for various values of CV, power, and product differences is taken from their publication.

Based on these tables, using log-transformed estimates of the parameters would result in a sample size estimate of 38 for a power of 0.8, ratio of 0.9, and CV = 0.20. If the assumed ratio is 1.1, the sample size is estimated as 32.

Equation (6.13) can also be used to approximate these sample sizes using log values for V and δ: $n = (1.69 + 0.85)^2[0.20/(0.223 - 0.105)]^2 = 19$ per sequence or 38 subjects in total, where 0.223 is the log of 1.25 and 0.105 is the absolute value of the log of 0.9.

Table 6.5 Sample Sizes for Given CV Power and Ratio (T/R) for Log-Transformed Parameters[a]

CV (%)	Power (%)	μ_r, μ_x 0.85	0.90	0.95	1.00	1.05	1.10	1.15	1.20
5.0	70	10	6	4	4	4	4	6	16
7.5		16	6	6	4	6	6	10	34
10.0		28	10	6	6	6	8	16	58
12.5		42	14	8	8	8	12	24	90
15.0		60	18	10	10	10	16	32	128
17.5		80	22	12	12	12	20	44	172
20.0		102	30	16	14	16	26	56	224
22.5		128	36	20	16	20	30	70	282
25.0		158	44	24	20	22	38	84	344
27.5		190	52	28	24	26	44	102	414
30.0		224	60	32	28	32	52	120	490
3.0	80	12	6	4	4	4	6	8	22
7.5		22	8	6	6	6	8	12	44
10.0		36	12	8	6	8	10	20	76
12.5		54	16	10	8	10	14	30	118
15.0		78	22	12	10	12	20	42	168
17.5		104	30	16	14	16	26	56	226
20.0		134	38	20	16	18	32	72	294
22.5		168	46	24	20	24	40	90	368
25.0		206	56	28	24	28	48	110	452
27.5		248	68	34	28	34	58	132	544
30.0		292	80	40	32	38	68	156	642
5.0	90	14	6	4	4	4	6	8	28
7.5		28	10	6	6	6	8	16	60
10.0		48	14	8	8	8	14	26	104
12.5		74	22	12	10	12	18	40	162
15.0		106	30	16	12	16	26	58	232
17.5		142	40	20	16	20	34	76	312
20.0		186	50	26	20	24	44	100	406
22.5		232	64	32	24	30	54	124	510
25.0		284	78	38	28	36	66	152	626
27.5		342	92	44	34	44	78	182	752
30.0		404	108	52	40	52	92	214	888

[a] *Source*: From Ref. [8].

For $\delta = 1.10$ ($\log = 0.0953$), the sample size is: $n = (1.69 + 0.85)^2[0.20/\ (0.223 - 0.0953)]^2 = 16$ per sequence or 32 subjects in total.

If the difference between products is specified as zero (ratio = 1.0), the value for $t_{\beta,\,2n-2}$ in Eq. (6.3) should be two sided (Table 6.2). For example, for 80% power (and a large sample size) use 1.28 rather than 0.84. In the example above with a ratio of 1.0 (0 difference between products), a power of 0.8, and a CV = 0.2, use a value of (approximately) 1.34 for $t_{\beta,\,2n-2}$.

$n = (1.75 + 1.34)^2[0.2/0.223]^2 = 7.7$ per group or 16 total subjects.

An Excel program to calculate the number of subjects required for a crossover study under various conditions of power and product differences, for both parametric and binary (binomial) data, is available on the disk accompanying this volume.

This approach to sample-size determination can also be used for studies where the outcome is dichotomous, often used as the criterion in clinical studies of bioequivalence (cured or not cured) for topically unabsorbed products or unabsorbed oral products such as sucralfate. This topic is presented in section 11.4.8.

KEY TERMS

Alpha level
Attribute
Beta error
Confidence interval
Delta
Power
Power curve
"Practical" significance
Sample size
Sampling plan
Sensitivity
Z transformation

EXERCISES

1. Two diets are to be compared with regard to weight gain of weanling rats. If the weight gain due to the diets differs by 10 g or more, we would like to be 80% sure that we obtain a significant result. How many rats should be in each group if the s.d. is estimated to be 5 and the test is performed at the 5% level?

2. How many rats per group would you use if the standard deviation were known to be equal to 5 in Problem 1?

3. In Example 3 where two antibiotics are being compared, how many patients would be needed for a study with $\alpha = 0.05$, $\beta = 0.10$, using a parallel design, and assuming that the new product must have a cure rate of 90% to be acceptable as a better product than the standard? (Cure rate for standard = 80%).

4. It is hypothesized that the difference between two drugs with regard to success rate is 0 (i.e., the drugs are not different). What size sample is needed to show a difference of 20% significant at the 5% level with a β error of 10%? (Assume that the response rate is about 50% for both drugs, a *conservative* estimate.) The study is a two independent samples design (parallel groups).

5. How many observations would be needed to estimate a response rate of about 50% within $\pm$ 15% (95% confidence limits)? How many observations would be needed to estimate a response rate of 20 $\pm$ 15%?

6. Your boss tells you to make a new tablet formulation that should have a dissolution time (90% dissolution) of 30 minutes. The previous formulation took 40 minutes to 90% dissolution. She tells you that she wants an α level of 5% and that if the new formulation really has a dissolution time of 30 minutes or less, she wants to be 99% sure that the statistical comparison will show significance. (This means that the β error is 1%.) The s.d. is approximately 10. What size sample would you use to test the new formulation?

7. In a clinical study comparing the effect of two drugs on blood pressure, 20 patients were to be tested on each drug (two groups). The change in blood pressure from baseline measurements was to be determined. The s.d., measured as the difference among individuals' responses, is *estimated* from past experience to be 5.
   (a) If the statistical test is done at the 5% level, what is the power of the test against an alternative of 3 mm Hg difference between the drugs ($H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$). This means: What is the probability that the test will show significance if the true difference between the drugs is 3 mm Hg or more ($H_a : \mu_1 - \mu_2 = 3$)?
   (b) What is the power if there are 50 people per group? α is 5%.

8. A tablet is produced with a labeled potency of 100 mg. The standard deviation is known to be 10. What size sample should be assayed if we want to have 90% power to detect a difference of 3 mg from the target? The test is done at the 5% level.

9. In a bioequivalence study, the ratio of AUCs is to be compared. A sample size of 12 subjects is used in a paired design. The standard deviation resulting from the statistical test is 0.25. What is the power of this test against a 20% difference if α is equal to 0.05?

10. How many samples would be needed to have 95% power for Example 8?

11. In a bioequivalence study, the maximum blood level is to be compared for two drugs. This is a crossover study (paired design) where each subject takes both drugs. Eighteen subjects entered the study with the following results. The observed difference is 10 μg/mL. The s.d. (from this experiment) is 40. A practical difference is considered to be 15 μg/mL. What is the power of the test for a 15-μg/mL difference for a two-sided test at the 5% level?
12. How many observations would you need to estimate a proportion within ±5% (95% confidence interval) if the expected proportion is 10%?
13. A parallel design is used to measure the effectiveness of a new antihypertensive drug. One group of patients receives the drug and the other group receives placebo. A difference of 6 mm Hg is considered to be of practical significance. The standard deviation (difference from baseline) is unknown but is estimated as 5 based on some preliminary data. Alpha is set at 5% and β at 10%. How many patients should be used in each group?
14. From Table 6.3, find the number of samples needed to determine the difference between the dissolution of two formulations for $\alpha = 0.05$, $\beta = 0.10$, $S = 25$, for a "practical" difference of 25 (minutes).

REFERENCES

1. United States Pharmacopeia, 23rd rev, and National Formulary, 18th ed. Rockville, MD: USP Pharmacopeial Convention, Inc., 1995.
2. U.S. Department of Defense Military Standard. Military Sampling Procedures and Tables for Inspection by Attributes (MIL-STD-105E). Washington, DC: U.S. Government Printing Office, 1989.
3. Guenther WC. Sample size formulas for normal theory tests. Am Stat 1981; 35:243.
4. Fleiss J. Statistical Methods for Rates and Proportions, 2nd ed. New York: Wiley, 1981.
5. Dixon WJ, Massey FJ Jr. Introduction to Statistical Analysis, 3rd ed. New York: McGraw-Hill, 1969.
6. Phillips KE. Power of the two one-sided tests procedure in bioequivalence. J Pharmacokinet Biopharm 1991; 18:137.
7. Chow S-C, Liu J-P. Design and Analysis of Bioavailability and Bioequivalence Studies. New York: Marcel Dekker, 1992.
8. Dilletti E, Hauschke D, Steinijans VW. Sample size determination: extended tables for the multiplicative model and bioequivalence ranges of 0.9 to 1.11 and 0.7 to 1.43. Int J Clin Pharmacol Toxicol 1991; 29:1.