Chapter 5
Interval estimation and testing

5.1 INTRODUCTION

As stated earlier, the reason for taking a sample is to obtain information about the unknown characteristics of a population or process. Two types of population characteristics are of special interest in business: the mean of a variable, and the proportion of elements belonging to a category. After the sample has been taken, these are estimated by the corresponding sample mean and proportion. In the majority of studies in business, these estimates (there could be many of them in a typical marketing survey) are all that is required of the sample.

The key to good estimates lies in the *design* of the sample, and this takes place *before* the sample is actually selected. A well-designed sample, we have argued, should be: (a) *randomly selected* from the population or process of interest; and (b) *large enough* so that the estimates will have the desired degree of accuracy.

Interval estimates may be used in place of, or as a supplement to, the "point" estimates we have encountered up to now. Rather than state, after a sample is taken, that a population characteristic is estimated to be such-and-such a number, it may on occasion be more informative to state that the characteristic is estimated to be in such-and-such an interval. Not any arbitrary interval, obviously, will do. If we are to make interval estimates, we want assurance that our statement will be correct (that is, that the interval will contain the population characteristic) with a given probability. Such intervals are known as *confidence intervals* and are described in the following two sections.

The remainder of the chapter deals with *statistical tests.* A statistical test is a rule—a prescription, if you like—for deciding which of two statements concerning an unknown population characteristic is true. No decision rule (statistical or other) is infallible. The attractive feature of statistical tests is that they allow the desision-maker to *control* the probability of making an error judged (by the decision-maker) to be the more serious.

5.2 INTERVAL ESTIMATION

In our discussion so far, we have argued that it is reasonable to use the sample mean ($\bar{X}$) as an estimator of the population mean (μ) of a variable, and the sample proportion (R) as an estimator of the population proportion (π) of a category. When the sample is taken, the numerical values of $\bar{X}$ and R are the estimates of μ and π respectively.

We know well, of course, that in general it is unlikely these estimates will equal μ or π exactly. Although we may say "we estimate μ and π to be 10.3 and 0.12," we certainly do not mean we consider these to be necessarily the true values of μ and π.

Instead of saying, for example, "μ is $\bar{X}$," we could say "μ is *in the interval* from $\bar{X} - c$ to $\bar{X} + c$." This is an interval around $\bar{X}$, usually abbreviated as $\bar{X} \pm c$, with c to be specified. $\bar{X} \pm c$ is an *interval estimator* of μ, and we may prefer it to the ordinary estimator $\bar{X}$.

For any choice of c, the statement "μ is in the interval $\bar{X} \pm c$" is sometimes correct (that is, the interval contains μ), sometimes not. Forming an arbitrary interval is not at all difficult, but forming an interval *having a given probability of containing* μ is not easy.

An interval estimator of a population characteristic which can be said to contain the population characteristic with given probability is called a *confidence interval*, and the given probability the *confidence level*. Approximate confidence intervals for the population mean of a variable and for the population proportion of a category, applicable when the population and sample sizes are large, are described in the box that follows.

Note that the intervals (5.1) and (5.2) make c equal to $U_{\alpha/2}S_{\bar{X}}$ and $U_{\alpha/2}S_R$ respectively, and can be calculated once the sample observations are available. The probability $1-\alpha$ is specified in advance. It could be large (e.g., 99%, 90%) or small (e.g., 30%, 10%), as desired. We shall continue to interpret "large N and n" to mean "larger than 200 and 100 respectively," according to the rule of thumb of the previous chapter. One should bear in mind, however, that it is a rule of thumb; more accurately, we should be saying that the probability of the interval containing the population characteristic tends to $1 - \alpha$ as n and N get larger.

To show that the intervals (5.1) and (5.2) contain π and μ respectively with probability approximately $1 - \alpha$ we need a property of large samples described in the next section; the proof is also in that section.

Before commenting on the features of these intervals, let us illustrate how they are calculated.

If the population (N) and the sample size (n) are large, the probability is approximately $1 - \alpha$ that the interval

$$R \pm U_{\alpha/2} S_R \tag{5.1}$$

contains the population proportion (π). Likewise, the probability is approximately $1 - \alpha$ that the interval

$$\bar{X} \pm U_{\alpha/2} S_{\bar{X}} \tag{5.2}$$

contains the population mean (μ). These intervals are known as *$100(1-\alpha)$% confidence intervals* for π and μ respectively, and $(1-\alpha)$ is known as the *confidence level*. Similar confidence intervals for $N\pi$ and $N\mu$ are

$$N(R \pm U_{\alpha/2} S_R) \quad \text{and} \quad N(\bar{X} \pm U_{\alpha/2} S_{\bar{X}}) \tag{5.3}$$

In the above expressions, S^2 is the *sample variance* of the variable X:

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2, \tag{5.4}$$

and

$$S_{\bar{X}} = \sqrt{\frac{S^2}{n} \frac{N-n}{N-1}}, \quad S_R = \sqrt{\frac{R(1-R)}{n} \frac{N-n}{N-1}}. \tag{5.5}$$

$U_{\alpha/2}$ for selected $(1-\alpha)$ are given in Table 5.1.

Table 5.1
$U_{\alpha/2}$ for selected $1 - \alpha$

$1-\alpha$	$U_{\alpha/2}$	$1-\alpha$	$U_{\alpha/2}$
0.99	2.576	0.80	1.282
0.95	1.960	0.60	0.842
0.90	1.645	0.50	0.674

Example 5.1 A random sample of $n = 500$ households from a population

of $N = 100,000$ was taken. 42% of the sampled households said they would buy an experimental product. We calculate

$$S_R = \sqrt{\frac{(0.42)(1-0.42)}{500}\frac{100000-500}{100000-1}} = 0.022.$$

Suppose that a 90% confidence interval is desired for the proportion of households in the population who intend to buy. Then, $1-\alpha = 0.90$, and $U_{\alpha/2} = 1.645$. Thus, $U_{\alpha/2}S_R = (1.645)(0.022) = 0.036$. The desired interval is from $(0.420 - 0.036)$ to $(0.420 + 0.036)$, or from 38.4% to 45.6%.

A 90% confidence interval for the number of households who intend to buy is (100,000)(0.420 ± 0.036), or from about 38,400 to 45,600.

Example 5.2 A random sample of $n = 800$ households in a town revealed that the average weekly household expenditure on food was $\bar{X} = \$95$ with a standard deviation of $S = 34$. (Just in case the meaning of S is not clear, 34 is the standard deviation of the 800 responses to the question, "How much is your household's weekly food expenditure?" The average of these 800 responses is, of course, 95.) There are $N = 10,000$ households in the town. A 95% confidence interval for the average weekly food expenditure by *all* households in the town would be given by

$$(95) \pm (1.96)\sqrt{\frac{(34)^2}{800}\frac{10000-800}{10000-1}}.$$

This is the interval from $(95 - 2.26)$ to $(95 + 2.26)$, or from \$92.74 to \$97.26.

A 95% confidence interval for the total household expenditures on food in the city is (10,000)(95 ± 2.26), or from about \$927,400 to \$972,600.

One should not interpret a *calculated* confidence interval as containing the population characteristic with the stated probability. For instance, it is not correct to say in Example 5.1 that the probability is 90% that the population proportion is between 38.4 and 45.6%. The population proportion is a given number, and either lies in that interval or does not—we do not know which is true. It is the *procedure* by which the intervals are calculated which is correct with probability 90%.

To understand this more clearly, imagine it is possible to select a large number of samples, all of the same size n, from the given population. After each sample is taken, imagine calculating the interval $\bar{X} \pm U_{\alpha/2}\ S_{\bar{X}}$ and stating that μ lies in this interval. Because $\bar{X}$ and $S_{\bar{X}}$ will vary from sample to sample, the location and width of the intervals will vary. See Figure 5.1.

The claim of Equation (5.2) is that, in the long run, $100(1-\alpha)\%$ of these intervals will contain (will "bracket") μ. (In Figure 5.1, all but one of

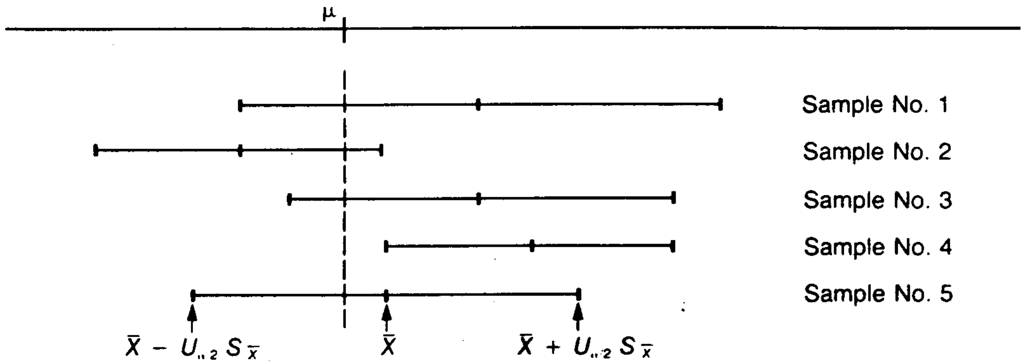

Figure 5.1
Interval estimates

the five confidence intervals shown—the exception being that from Sample No. 4—contain μ.) The statement that the population mean lies in the interval $\bar{X} \pm U_{\alpha/2} S_{\bar{X}}$ will therefore be correct in $100(1-\alpha)\%$ of the samples in the long run.

The calculated confidence interval, however, conveys some information about the reliability of the sample estimate. The intervals (5.1) and (5.2), it will be noted, are formed around the estimate of the population characteristic, and their width depends on the size of the sample, the variability of the observations, and the desired confidence level.

As an illustration, consider the width of the confidence interval for a population mean: it is approximately twice the amount

$$c = U_{\alpha/2}\sqrt{S^2(\frac{1}{n} - \frac{1}{N})}.$$

The smaller the confidence level, $1 - \alpha$, the smaller is $U_{\alpha/2}$. Other things being equal, therefore, the larger the sample size, the smaller the variability of the sample observations (as measured by S^2), and the lower the confidence level, the narrower the interval—which is as one intuitively expects.

The reader should always bear in mind that a confidence interval is *not* a *necessary* appendage to the sample estimate. In most business samples designed to obtain estimates of a large number of population characteristics, a routine reporting of all possible confidence intervals would most certainly confuse rather than enlighten.*

* In addition to the confidence intervals based on large samples described in this section, it is possible to construct a confidence interval for a popula-

5.3 FURTHER PROPERTIES OF LARGE SAMPLES

In order to understand the derivation of the confidence intervals in the previous section, and of the statistical tests described in the next section, we must state and briefly explain two more properties of large samples.

> If the population (N) and the sample (n) are large, the probability distributions of the random variables
>
> $$U_1 = \frac{\bar{X} - \mu}{S_{\bar{X}}}, \quad U_2 = \frac{R - \pi}{S_R}, \tag{5.6}$$
>
> are approximately standard normal.

As usual, $\bar{X}$ is the sample mean and μ the population mean of a variable X, while R is the sample and π the population relative frequency of a category. $S^2_{\bar{X}}$, defined in Equations (5.5), is an estimator of the variance of $\bar{X}$,

$$Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}, \tag{5.7}$$

obtained by replacing the population variance (σ^2) with the sample variance (S^2). Similarly, S^2_R, also defined in Equations (5.5), is an estimator of the variance of R,

$$Var(R) = \frac{\pi(1-\pi)}{n}\frac{N-n}{N-1}, \tag{5.8}$$

obtained by replacing the population proportion π by the sample proportion R.

Each U is a "standardized" random variable, measuring the difference between $\bar{X}$ (or R) and its expected value μ (or π), with the difference expressed in estimated standard deviations of its distribution. Because $\bar{X}$, R, $S_{\bar{X}}$, and S_R will vary from sample to sample, so will the U's, which are functions of these random variables.

Consider Example 4.1 of Chapter 4 by way of illustration. In this case, $N = 10$, $\pi = 0.7$, and $\mu = 0.9$. Refer to Table 4.1 of Chapter 4 and suppose

---

tion proportion in a sample of *any* size with or without replacement. Similar, "any-sample-size" confidence intervals for the population mean of a variable, however, can be formed only in the case where the population distribution of the variable is of a special form (for example, normal, Poisson, etc.) and the sample is with replacement.

that Outcome No. 3 occurs, that is, $X_1 = 0$, $X_2 = 0$, and $X_3 = 2$. It follows that $\bar{X} = 2/3 = 0.667$, $R = 1/3 = 0.333$, and the sample variance is

$$S^2 = \frac{1}{n}\sum(X_i - \bar{X})^2 = \frac{1}{3}[(0-0.667)^2 + (0-0.667)^2 + (2-0.667)^2] = 0.889.$$

Therefore,

$$S_{\bar{X}} = \sqrt{\frac{0.889}{3}\frac{10-3}{10-1}} = 0.480, \qquad S_R = \sqrt{\frac{(0.333)(1-0.333)}{3}\frac{10-3}{10-1}} = 0.240,$$

$$U_1 = \frac{0.667 - 0.9}{0.480} = -0.485,$$

and

$$U_2 = \frac{0.333 - 0.7}{0.240} = -1.529.$$

For example, $U_1 = -0.485$ indicates that the observed value of the sample mean lies 0.485 estimated standard deviations to the left of its expected value, and $U_2 = -1.529$ shows that the observed value of R lies 1.529 estimated standard deviations to the left of the mean of its distribution.

For each sample outcome listed in Table 4.1 of Chapter 4, there corresponds a value of U_1 and U_2. The U's are random variables, and their probability distributions can be constructed in the usual way. We shall not do so, however, because both N and n are small, and no useful purpose will be served by continuing the calculations.

Consider now Figure 5.2. The histograms depict the probability distributions of U_1 for samples of size 3, 10, 20, and 30 *with* replacement drawn from the population of Example 4.1 of Chapter 4. Each of these probability distributions was routinely constructed from a list of all possible sample outcomes, their probabilities, and the associated values of U_1; a computer program was used to perform the long and tedious numerical calculations. When $S^2 = 0$, which occurs when all the sample observations are identical, U_1 has an indeterminate value; the probability of an indeterminate value is indicated by a separate bar in the histograms of Figure 5.2. Note that the probability of an indeterminate value decreases as the sample size increases. Compare the histograms with the superimposed standard normal distribution. As the sample size increases, the standard normal distribution provides an increasingly better approximation to the actual distribution of U_1.

What is illustrated for U_1 and sampling with replacement can be shown to hold for U_2 and for sampling without replacement: for large N and n ($n < N$), the actual probability distribution of any U_i may be approximated by the standard normal distribution.

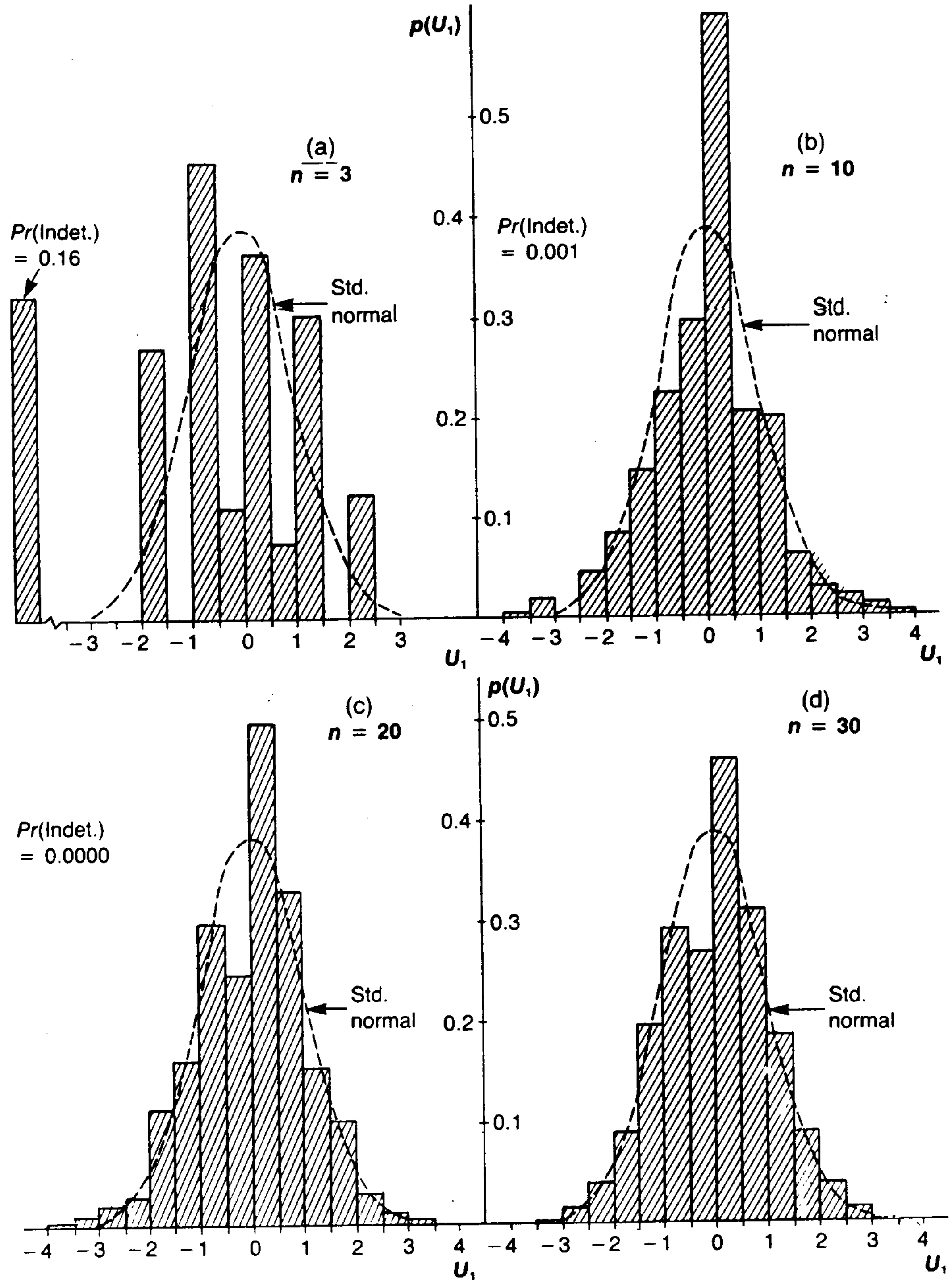

Figure 5.2
Distribution of U_1

It has not been made clear so far just *how large* N and n must be for the normal approximations to be satisfactory. Unfortunately, no exact guidance can be given. In statistical theory, the large-sample properties are proven for *infinitely* large N and n. Empirical investigations, however, show the approximation to be surprisingly good in some cases for samples as low as 30 or smaller. On the other hand, there are other cases where the approximation is poor even for very large samples.

Roughly speaking, the more symmetric the population distribution of the variable X (in the case of U_1) or the closer π is to 0.5 (in the case of U_2), the smaller is the sample size required.

As a rule of thumb, the reader would probably be safe in assuming that for $n \geq 100$ and $N \geq 200$—conditions easily satisfied in most samples used in business—the normal approximation is satisfactory, and that the results based on this approximation (to be described in the following sections) are applicable.

We would now like to show that the probability that the interval (5.2) will contain ("bracket") μ is approximately $1 - \alpha$. The proof concerning the interval (5.1) is very similar, and is left as an exercise for the reader.

¶ For large N and n, the distribution of the ratio $U_1 = (\bar{X} - \mu)/S_{\bar{X}}$ is approximately the standard normal. Let $U_{\alpha/2}$ be a number such that the probability that the standard normal variable will exceed that number is $\alpha/2$. By the symmetry of the normal distribution, $Pr(-U_{\alpha/2} \leq U_1 \leq U_{\alpha/2}) = 1 - \alpha$. Substituting $(\bar{X} - \mu)/S_R$ for U_1, we have

$$Pr(-U_{\alpha/2} \leq \frac{\bar{X} - \mu}{S_{\bar{X}}} \leq U_{\alpha/2}) = 1 - \alpha. \tag{5.9}$$

Consider the expression within the parentheses. We apply the same two rules concerning inequalities stated in Chapter 2.* Multiplying all three terms by (the positive) $S_{\bar{X}}$, we get $-U_{\alpha/2}S_{\bar{X}} \leq \bar{X} - \mu \leq U_{\alpha/2}S_{\bar{X}}$. Multiplying these terms by -1, we reverse the inequalities and get $U_{\alpha/2}S_{\bar{X}} \geq \mu - \bar{X} \geq -U_{\alpha/2}S_{\bar{X}}$. Adding $\bar{X}$ to all three terms gives $\bar{X} - U_{\alpha/2}S_{\bar{X}} \leq \mu \leq \bar{X} + U_{\alpha/2}S_{\bar{X}}$. The first expression implies the last. All we have done is write the original inequalities in a different form. Therefore, Equation (5.9) implies that

$$Pr(\bar{X} - U_{\alpha/2}S_{\bar{X}} \leq \mu \leq \bar{X} + U_{\alpha/2}S_{\bar{X}}) = 1 - \alpha. \tag{5.10}$$

In words, the probability is $1 - \alpha$ that the interval from $(\bar{X} - U_{\alpha/2}S_{\bar{X}})$ to $(\bar{X} + U_{\alpha/2}S_{\bar{X}})$ contains the population mean. This interval is therefore a $100(1 - \alpha)\%$ confidence interval for μ, and the proof is complete. ¶

It may be noted that the type of confidence interval described in these sections (two-sided, symmetric, and centered around the sample estimate) is

* If $a \leq X \leq b$ then $ca \leq cX \leq cb$ for $c > 0$, or $ca \geq cX \geq cb$ for $c < 0$. Also, if $a \leq X \leq b$ then $a + c \leq X + c \leq b + c$ for any c.

by no means unique. Instead of an interval of the form $\bar{X}-c \leq \mu \leq \bar{X}+c$, one can construct confidence intervals of the form $\mu \leq c$, $\mu \geq c$, or $c_1 \leq \mu \leq c_2$ (where c, c_1, and c_2 depend on the sample observations), all having the same probability of containing μ.

For example, we may begin with the observation that the following statements are approximately true when the random variable U has the unit normal distribution:

$$\begin{aligned}
&Pr(-1.28 \leq U \leq 1.28) = Pr(-1.04 \leq U \leq 1.64) = \\
&\quad = Pr(U \leq 0.84) = Pr(U \geq -0.84) = 0.80.
\end{aligned}$$

Substituting $(\bar{X}-\mu)/S_{\bar{X}}$ for U and following the same approach as above, we find that:

$$\begin{aligned}
&Pr(\bar{X} - 1.28S_{\bar{X}} \leq \mu \leq \bar{X} + 1.28S_{\bar{X}}) = \\
&\quad = Pr(\bar{X} - 1.64S_{\bar{X}} \leq \mu \leq \bar{X} + 1.04S_{\bar{X}}) = \\
&\quad = Pr(\mu \leq \bar{X} + 0.84S_{\bar{X}}) = \\
&\quad = Pr(\mu \geq \bar{X} - 0.84S_{\bar{X}}) = 0.80.
\end{aligned}$$

Therefore, the intervals:

(a) from $\bar{X} - 1.28S_{\bar{X}}$ to $\bar{X} + 1.28S_{\bar{X}}$,
(b) from $\bar{X} - 1.64S_{\bar{X}}$ to $\bar{X} + 1.04S_{\bar{X}}$,
(c) from $\bar{X} - 0.84S_{\bar{X}}$ to $+\infty$,
(d) from $-\infty$ to $\bar{X} + 0.84S_{\bar{X}}$

are all approximate 80% confidence intervals for μ. An infinite number of type (b) intervals can be constructed by varying appropriately the limits of the probability statement. The symmetric interval (a) is preferable to any asymmetric one of type (b) because it is narrower: in our example, the width of (a) is $(2)(1.28)S_{\bar{X}} = 2.56S_{\bar{X}}$, while that of (b) is $1.64S_{\bar{X}} + 1.04S_{\bar{X}} = 2.68S_{\bar{X}}$. Due to a property of the standard normal distribution, the same is true for all asymmetric intervals (b). The choice between (a), (c), and (d), however, is not obvious and depends on the specific problem.

5.4 UNDERSTANDING STATISTICAL TESTS

In this section, we examine with some care two types of statistical tests. As will soon be clear, there are many such tests. It is not our intention to examine them all exhaustively, but to identify in the two tests of this section the features and problems that most tests have in common.

Example 5.3 A certain brand of light bulbs has a "rated life of 1,000 hours." This rating is assigned by the manufacturer. The small print on the package explains that the *average* life of this brand of bulbs is warranted to be more than 1,000 hours. (It is understood that the life of individual light bulbs varies; some of this variability could perhaps be removed by better quality control, but some is inherent in the manufacturing process and cannot be eliminated.)

A batch of 10,000 bulbs has been produced. Before it is shipped out, a test must be made to determine if the quality of the batch is consistent with the rating. At issue, therefore, is whether or not the average life of the bulbs in the batch is more than 1,000 hours. Measuring the life of all 10,000 bulbs is obviously out of the question since the measurement is destructive: life is measured by letting the bulb burn until it burns out. A sample must be used.

In one sense, the problem is simple. The manufacturer could select a random sample of light bulbs, measure their life duration, and calculate the average life of the bulbs in the sample. If this sample average is greater than 1,000 hours, the conclusion could be that the batch is Good (that is, the quality rating is justified); the batch would then be released. On the other hand, if the sample average is less than or equal to 1,000 hours, the conclusion could be that the batch is Bad (does not meet the quality standard) and its release would be withheld.

Let μ be the (unknown) average life of the bulbs *in the batch.* The issue then is whether or not μ is more than 1,000 hours. A statistician would say that the problem involves two *hypotheses* concerning μ:

$$\begin{array}{lllll} H_1: & \mu & \leq & 1000 & \quad\text{(Batch is Bad)} \\ H_2: & \mu & > & 1000 & \quad\text{(Batch is Good)} \end{array}$$

Obviously, if H_1 is true, H_2 is false, and vice versa.

A decision must be based on a sample. The previously suggested *decision rule* can be expressed as

$$\begin{array}{rl} \text{Accept } H_1 \text{ (i.e., conclude batch is Bad) if} & \bar{X} \leq 1000, \\ \text{Reject } H_1 \text{ (i.e., conclude batch is Good) if} & \bar{X} > 1000, \end{array} \tag{5.11}$$

where $\bar{X}$ is the average life of n randomly selected light bulbs. (The terms "accept" and "reject" are part of the established terminology but need not be taken literally. By "accept H_1" we mean "decide in favor of H_1," and by "reject H_1" "decide in favor of H_2.")

This decision rule may be perfectly sensible, but it is not at all flexible. Let us therefore make it a little more general, as follows:

$$\begin{array}{rl} \text{Accept } H_1 \text{ if} & \bar{X} \leq c, \\ \text{Reject } H_1 \text{ if} & \bar{X} > c, \end{array} \tag{5.12}$$

Table 5.2
Hypotheses, actions, and consequences

	Acts	
Events	Accept H_1	Reject H_1
H_1 is true	No error	Type I error
H_1 is false	Type II error	No error

where c is a number to be determined (as a special case, it *could* be made equal to 1,000). In this version, the decision rule recommends accepting H_1 when $\bar{X}$ is "small," and rejecting H_1 when it is "large"—with c distinguishing small from large values of $\bar{X}$.

Table 5.2 shows the two hypotheses, the two possible decisions, and the associated consequences. If H_1 is true and is accepted, or false and rejected, no error is made. But if H_1 is true and is rejected, or false and accepted, an error is made. These are two different kinds of error, and are distinguished in Table 5.2 as errors of Type I and II.

In general, by a Type I error we shall understand that of rejecting H_1 when H_1 is true—whatever the meaning of H_1 happens to be. In our example, a Type I error means that a Bad batch is declared Good. Possible consequences include customer complaints, an erosion of the manufacturer's quality image, etc.

A Type II error, on the other hand, is that of accepting H_1 when it is false—again, whatever the meaning of H_1 may be. In the present example, a Type II error means that a Good batch is declared Bad. It could result in the lot being scrapped or reworked.

The ideal decision rule should guarantee that the probabilities of both types of error are zero. A little thought will lead one to the conclusion that this is impossible, unless the sample is the entire population. It is possible, however, to formulate a rule guaranteeing that the probability of *one* type of error—whichever is the more serious—does not exceed a given number.

Let us suppose that a Type I error (Bad lot declared Good) is considered more serious, and that the probability of such an error should not exceed 20%. The problem, therefore, is to find a decision rule ensuring that the probability of a Type I error does not exceed 20%.

Our problem is in fact a special case of that described in the box that follows, and a decision rule having the required property is given by (5.14).

Before explaining the derivation of this decision rule, let us illustrate its implementation.

In our problem, $\mu_0 = 1,000$ and $\alpha = 0.20$. Suppose that a random sample of $n = 100$ light bulbs is taken, that the average life of the bulbs in the sample is $\bar{X} = 1,010$ hours, and their standard deviation $S = 80$.

When N and n are large, the approximate decision rule for testing

$$\begin{aligned} H_1 &: \mu \leq \mu_0, \\ H_2 &: \mu > \mu_0, \end{aligned} \tag{5.13}$$

μ_0 a given number, so that the probability of Type I error does not exceed α and that of Type II error does not exceed $1 - \alpha$, is to

$$\begin{aligned} \text{Accept } H_1 \text{ if } \quad \bar{X} &\leq \mu_0 + U_\alpha S_{\bar{X}}, \\ \text{Reject } H_1 \text{ if } \quad \bar{X} &> \mu_0 + U_\alpha S_{\bar{X}}. \end{aligned} \tag{5.14}$$

Selected values of U_α are given in Table 5.3. $S_{\bar{X}}$ is given in Equations (5.5).

Calculate

$$S_{\bar{X}} = \sqrt{\frac{(80)^2}{100} \frac{10000 - 100}{10000 - 1}} = 7.96.$$

$U_\alpha = U_{0.20} = 0.842$, and $\mu_0 + U_\alpha S_{\bar{X}} = (1000) + (0.842)(7.96) = 1{,}006.70$.

Since $\bar{X} = 1010 > 1006.7$, the decision rule (5.14) requires that H_1 be rejected—in other words, that the batch be declared Good.

It is clear that the decision rule (5.14) implies a c in (5.12) equal to $\mu_0 + U_\alpha S_{\bar{X}}$. Note that when $\alpha = 0.50$ (a value appropriate when the two types of error are considered equally serious), $U_\alpha = 0$, and the rule becomes: Accept H_1 if $\bar{X} < \mu_0$. In our example, the decision rule (5.12) is therefore appropriate in just such a situation.

¶ We would now like to show that the decision rule (5.14) is indeed such that the probability of Type I error does not exceed α, and that of Type II error does not exceed $1 - \alpha$.

By definition, a Type I error occurs when $\mu \leq \mu_0$, and a Type II error when $\mu > \mu_0$. The proof, then, is in two steps, as follows.

First, suppose that the true value of μ is less than or equal to μ_0, i.e., H_1 is true. The probability of a Type I error is the probability of rejecting H_1, i.e.,

$$\begin{aligned} Pr(\bar{X} > \mu_0 + U_\alpha S_{\bar{X}}) &= Pr\Big(\frac{\bar{X} - \mu_0}{S_{\bar{X}}} > U_\alpha\Big) \\ &= Pr\Big(\frac{\bar{X} - \mu + \mu - \mu_0}{S_{\bar{X}}} > U_\alpha\Big) \\ &= Pr\Big(\frac{\bar{X} - \mu}{S_{\bar{X}}} > U_\alpha - \frac{\mu - \mu_0}{S_{\bar{X}}}\Big) \\ &= Pr(U_1 > U_\alpha - e) \\ &= Pr(U_1 > U''), \end{aligned}$$

where U_1, for large N and n, has approximately the standard normal distribution and e is less than or equal to zero. Therefore, $U'' \geq U_\alpha$; as illustrated in Figure 5.3, $Pr(U_1 > U'') \leq \alpha$. We conclude that the probability of a Type I error does not exceed α.

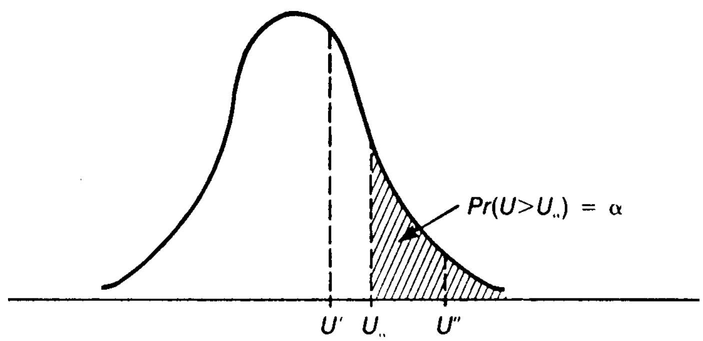

Figure 5.3
Standard normal distribution

For the second step of the proof, suppose that μ is greater than μ_0, i.e., H_1 is false. The probability of Type II error is the probability of accepting H_1, or

$$\begin{aligned} Pr(\bar{X} \leq \mu_0 + U_\alpha S_{\bar{X}}) &= Pr\Big(\frac{\bar{X} - \mu_0}{S_{\bar{X}}} \leq U_\alpha\Big) \\ &= Pr\Big(\frac{\bar{X} - \mu + \mu - \mu_0}{S_{\bar{X}}} \leq U_\alpha\Big) \\ &= Pr\Big(\frac{\bar{X} - \mu}{S_{\bar{X}}} \leq U_\alpha - \frac{\mu - \mu_0}{S_{\bar{X}}}\Big) \\ &= Pr(U_1 \leq U_\alpha - e) \\ &= Pr(U_1 \leq U') \end{aligned}$$

where e is positive and $U' < U_\alpha$. As illustrated in Figure 5.3, $Pr(U_1 \leq U') < 1-\alpha$, that is, the probability of Type II error cannot exceed $1 - \alpha$, and the proof is complete. ¶

The choice of α, hence also of U_α, should depend on the relative seriousness of the two types of error. The more serious the consequences of a Type I error in relation to those of a Type II error, the lower α should be, and vice versa. α can be any number between 0 and 1.

If it is desired to restrict the probability of a Type II error in a given situation to be, say, no more than 0.10, this can be accomplished by setting $\alpha = 1 - 0.10 = 0.90$. *If the two errors are equally serious*, set $\alpha = 0.50$, in which case the decision rule is simplified as we have already seen.

A final note. The $\leq$ in the statement of H_1 appears to suggest that this hypothesis is somewhat "heavier" than H_2. In fact, the case $\mu = \mu_0$, say, $\mu = 1$ means *precise* equality (e.g., $\mu = 1$ means precisely 1, and not 0.999999 or some such number close to 1). Unless it happens that μ is discrete, it makes no difference which hypothesis the case $\mu = \mu_0$ is assigned to or whether it is omitted entirely.

A similar test concerning a population proportion can be developed along the same lines, and is summarized by expressions (5.15) and (5.16) in the box that follows.

For large N and n, the approximate decision rule for testing

$$\begin{aligned} H_1 &: \pi \leq \pi_0, \\ H_2 &: \pi > \pi_0, \end{aligned} \tag{5.15}$$

(π_0 a given number) so that the probability of Type I error does not exceed α and that of Type II error does not exceed $1 - \alpha$ is to

$$\begin{aligned} \text{Accept } H_1 \text{ if } \quad & R \leq \pi_0 + U_\alpha S_R \\ \text{Reject } H_1 \text{ if } \quad & R > \pi_0 + U_\alpha S_R. \end{aligned} \tag{5.16}$$

Selected values of U_α are given in Table 5.3. S_R is given in Equations (5.5).

Example 5.4 Suppose that by a "rating of 1,000 hours," the light bulb manufacturer means "relatively few bulbs last less than 1,000 hours." (Many quality standards in manufacturing have this type of interpretation.) More specifically, suppose that "relatively few" means "5% or less." Let π represent the proportion of light bulbs in the batch which last less than 1,000 hours. We wish to test

$$H_1 : \pi \leq 0.05 \qquad \text{(Batch is Good)},$$
$$H_2 : \pi > 0.05 \qquad \text{(Batch is Bad)}.$$

In this case, a Type I error occurs when a Good lot is declared Bad, and a Type II error when a Bad lot is declared Good. Suppose it is the latter error which is considered more serious, and that the probability of Type II error should not exceed 10%. We set $\alpha = 1 - 0.10 = 0.90$, which implies $U_\alpha = -1.282$.

A random sample without replacement of size $n = 100$ is taken from the batch of $N = 10{,}000$ light bulbs. Four of the 100 bulbs ($R = 0.04$) are found to last less than 1,000 hours. We calculate

$$S_R = \sqrt{\frac{(0.04)(1-0.04)}{100}\frac{10000-100}{10000-1}} = 0.019,$$

and then $\pi_0 + U_\alpha S_R = (0.050) - (1.282)(0.019) = 0.026$. Since $R > 0.026$, H_1 is rejected and the batch is declared Bad.

This somewhat surprising conclusion is dictated by the choice of α which guards against the occurrence of a Type II error. If the two errors were considered equally serious, or if the Type I error were thought to be more serious, the opposite conclusion would have been reached.

5.5 THE STRUCTURE OF STATISTICAL TESTS

The two problems described in the previous section have several features in common.

There are two possible acts (release or withhold the batch). The consequences of these acts depend on the unknown value of a population or process characteristic (μ or π). The possible values of this characteristic are partitioned into two mutually exclusive and exhaustive sets or intervals (e.g., $\mu \leq 1000$, $\mu > 1000$), so that one act is optimal if the characteristic is in the first interval, and the other is optimal if it is in the second one. No information is available concerning the likelihood of the possible values of the population characteristic; a sample is taken to obtain some such information. There is no additional information concerning the consequences of the two acts beyond the fact that each is optimal for certain values of the population characteristic, and that perhaps some consequences are more serious than others.

These are indeed the common features of situations for which most statistical tests are designed.

Let us call the population characteristic θ (it could be μ, π, or something else), and the two sets or intervals S_1 and S_2. Denote the two possible acts

a_1 and a_2; assume a_1 is preferable if θ is in S_1, and a_2 is preferable if θ is in S_2.

Attention, therefore, focusses on two questions: Is θ in S_1? Is it in S_2? These questions correspond to the hypotheses: $H_1 : \theta\ in\ S_1$, and $H_2 : \theta\ in\ S_2$. The two acts can be labelled a_1 : *Accept* H_1, and a_2 : *Reject* H_1. And, since the consequences cannot be made more precise, they are indicated broadly as "No error," "Type I error," "Type II error," as shown in Table 5.4.

Table 5.4
Events, acts, and consequences

	Acts	
Events	a_1 : Accept H_1	a_2 : Reject H_1
$H_1 : \theta$ in S_1	No error	Type I error
$H_2 : \theta$ in S_2	Type II error	No error

A test is a decision rule prescribing when to choose a_1 and when a_2. Clearly, there are many possible decision rules for any given hypotheses. At a minimum, it is reasonable to require that the rule be based on the sample (that is, not to be arbitrary as, for instance, the rule: Reject H_1 if the moon is full), and that it allow control of the probabilities of the two types of error.

All the tests of this chapter have the property that the probability of a Type I error will not exceed α, and that of a Type II error will not exceed $1 - \alpha$, no matter what the value of θ happens to be. (Such a test, however, may not be unique; that is, there may be several tests with this property for given hypotheses.)

There are, obviously, many possible hypotheses, and, consequently, many statistical tests. There are, to begin with, many population characteristics which conceivably could matter in a given situation (e.g., the median, the mode, or the variance of a variable, to name just three). There are many types of hypotheses which can be formulated (one pair that comes quickly to mind: $H_1 : \mu_1 \leq \mu \leq \mu_2$, $H_2 : \mu > \mu_2$ or $\mu < \mu_1$, where μ_1 and μ_2 are given numbers). It is possible to formulate hypotheses involving jointly two or more population characteristics (e.g., that the population mean and variance are in a specified region), or more than two sets or intervals of values of a population characteristic (although multiple-action problems rapidly become quite complicated). And, finally, there may be several ways of testing the given hypotheses.

Not all hypotheses are testable, however, and not all tests are designed with business requirements in mind. Consider, for example, the oft-cited hypotheses $H_1 : \mu = \mu_0$ *vs.* $H_2 : \mu \neq \mu_0$. μ is the population mean of a variable, and μ_0 a given number—say, 10. A test of $H_1 : \mu = 10$ is precisely that—a test of the hypothesis that μ equals 10 *precisely* (not 10.00000001, or 9.9999999, or some other such number). Unless μ takes integer values only, it is very unlikely in typical business problems that the unknown population mean of a variable equals a given number exactly, and thus we know in advance that H_1 is false. H_1 will be rejected if the sample size is large enough. The non-rejection of H_1 is simply a consequence of a small sample. (This shortcoming is shared by some of the more frequently quoted tests, as will soon be explained.)

In the remainder of this chapter, we describe (but do not derive) a number of frequently cited tests. Their implementation—despite a first appearance of complexity—is quite easy. When implementing a test, however, it is worthwhile to heed three broad admonitions.

- Two-action tests are designed so that the probability of a Type I error does not exceed α, and that of a Type II error does not exceed $1 - \alpha$. A Type I error, by definition, is the rejection of H_1 when it is true; a Type II error is the acceptance of H_1 when it is false. *The real meaning and consequences of these errors, however, depend on the situation.* When applying a given test, therefore, one should understand what these errors imply and choose α or $1 - \alpha$ appropriately. There is an unfortunate tendency in practice (occasionally reinforced by the manner in which tables of critical values are presented) to select a low value of α automatically and indiscriminately, without regard to the situation.
- With the exception of Section 5.6, *all the tests of this chapter are approximate.* Although the calculations are carried out to three decimal places in order to avoid ambiguity, in practice one should not lose sight of the fact that the conditions "large N and n," as well as the value of α, are not and cannot be specified precisely.
- A statistical test is a device for *controlling* the probability of one type of error, and not—as many erroneously believe—a means of proving or disproving one or the other hypothesis. It is quite possible that two parties, presented with the same information, will reach diametrically opposite conclusions if they guard against different kinds of error.

5.6 INFERENCE FOR SPECIAL POPULATION DISTRIBUTIONS

When the form of the distribution of a variable in the population or process is mathematically tractable, it is sometimes possible to construct confidence intervals or tests that do not require the sample size to be large—they apply to small as well as large samples. Perhaps the simplest and most elegant results are obtained when this distribution is *normal.*

If the distribution of the measurement X of an *independent process* is normal with mean μ and standard deviation σ, the probability distribution of the ratio

$$U_1 = \frac{\bar{X} - \mu}{S_{\bar{X}}},$$

in samples of any size, n, is the t distribution with parameter $\nu = n - 1$.

The t distribution is defined in Appendix 2. It is symmetric, centered at 0, and looks much like the standard normal. As usual, $\bar{X}$ is the sample mean, and $S_{\bar{X}}$ is defined by Equations (5.5).

From the table of the t distribution in Appendix 4G, it is possible to determine a number—call it $T_{\alpha/2}$—such that $Pr(-T_{\alpha/2} \leq U_1 \leq T_{\alpha/2}) = 1-\alpha$. (For example, if $n = 5$ and $1-\alpha = 0.90$, then $\nu = 4$ and $T_{0.05} = 2.132$.) Proceeding exactly as in Section 5.3, we find that

$$Pr(\bar{X} - T_{\alpha/2}S_{\bar{X}} \leq \mu \leq \bar{X} + T_{\alpha/2}S_{\bar{X}}).$$

It follows that the interval

$$\bar{X} \pm T_{\alpha/2}S_{\bar{X}}$$

is a $100(1-\alpha)\%$ confidence interval for the process mean of measurement X, μ, *valid for any sample size.*

When $n > 30$, $T_{\alpha/2}$ is approximately equal to $U_{\alpha/2}$, and this special result merges with the more general ones of Section 5.2.

Following a procedure similar to that of Section 5.4, it is also fairly easy to show that the decision rule for testing $H_1 : \mu \leq \mu_o$ *vs.* $H_2 : \mu > \mu_o$, where μ is the mean of an independent normal process, is given by (5.14), except that U_α is replaced by T_α. T_α is a number such that the probability that a variable having the t distribution with parameter ν will exceed that number is α; see Appendix 4G.

The same results apply if the sample is random and with replacement from a finite population in which the distribution of the variable X is normal.

The reader should be warned against uncritically accepting the normal assumption and making inferences with small samples on the basis of this assumption alone. Because the normal distribution is mathematically tractable and lends itself to many elegant results, the tendency of textbooks and researchers alike is to give this distribution a more prominent role than it deserves: in business practice, few population or process distributions are normal. The test of the next section may be used to determine if the normality assumption is valid in a given case.

5.7 TESTS CONCERNING THE FORM OF THE POPULATION DISTRIBUTION

Suppose that the elements of a population or process are classified into s mutually exclusive and collectively exhaustive categories $C_1, C_2, \ldots, C_s$. We wish to test the hypothesis that the population relative frequencies are equal to specified numbers $\pi_{1o}, \pi_{2o}, \ldots, \pi_{so}$, against the alternative hypothesis that at least one of the π_i is not equal to the specified number.

Example 5.5 A casino is testing a die for fairness. The die is fair if the six faces show up with equal relative frequencies (π_i) in the long run.

$$\begin{array}{lll} H_1: & \pi_1 = 1/6,\ \pi_2 = 1/6,\ \ldots, \pi_6 = 1/6 & \text{(Die is fair)} \\ H_2: & \pi_i \neq 1/6 \ \text{ for at least one } \pi_i & \text{(Die not fair)} \end{array}$$

The die will be rolled a number of times, and the relative frequencies with which the six faces show up observed. How should one decide whether the die is fair or not?

An approximate decision rule for this situation is described in the following box.

For a large random sample with replacement or sample from an independent process, the approximate decision rule for testing

$$\begin{aligned} &H_1 : \pi_1 = \pi_{1o}, \pi_2 = \pi_{2o}, \ldots, \pi_s = \pi_{so}, \\ &H_2 : \text{At least one } \ \pi_i \neq \pi_{io}, \end{aligned} \tag{5.17}$$

so that the probability of a Type I error equals α and that of a Type II error does not exceed $1 - \alpha$, is to:

$$\begin{aligned} &\text{Accept } H_1 \text{ if } V \leq V_{\alpha;s-1}, \\ &\text{Reject } H_1 \text{ if } V > V_{\alpha;s-1}, \end{aligned} \tag{5.18}$$

where

$$V = n \sum_{i=1}^{s} \frac{(R_i - \pi_{io})^2}{\pi_{io}}. \tag{5.19}$$

The π_{io} are given numbers, R_i is the relative frequency of category C_i in the sample, n is the sample size, and $V_{\alpha;s-1}$ is as given in Appendix 4H.

Before illustrating the calculations, let us sketch the derivation of this test. If H_1 is true, that is, if the π_i equal π_{io}, the R_i will tend to be close to the π_{io}, and the value of V will tend to be close to zero. If, on the other hand, the true π_i deviate from the hypothesized values under H_1, π_{io}, the R_i will tend to deviate from the π_{io}, and V will tend to be greater than zero. We want to accept H_1 when V is small, and to reject it when V is large:

$$\begin{aligned} &\text{Accept } H_1 \text{ if } \quad V \leq c, \\ &\text{Reject } H_1 \text{ if } \quad V > c. \end{aligned}$$

c, the "critical value" of this test, distinguishes "small" from "large" V values. As always, we would like to determine c so that the probability of a Type I error does not exceed α, and that of a Type II error $1-\alpha$.

In mathematical statistics, it is shown that *if H_1 is true and the sample large and with replacement*, the probability distribution of V is approximately *chi-square* with parameter $\lambda = s-1$. (The definition of this distribution can be found in Appendix 2.) Denote by $V_{\alpha;s-1}$ the number such that the probability of a chi-square random variable with parameter $\lambda = s-1$ exceeding that number is α. These numbers are tabulated in Appendix 4H. It follows that if c is made equal to this number, the probability of a Type I error will not exceed α (in fact, will equal α). Intuitively, it should be clear that the probability of a Type II error is greatest when the π_i are very close to the π_{io}, at which point the probability of a Type II error is $1-\alpha$.

For calculations by hand, any one of the following versions of Equation (5.19) may be used:

$$V = \sum_{i=1}^{s} \frac{(F_i - n\pi_{io})^2}{n\pi_{io}} = n\Big(\frac{\sum_{i=1}^{s} R_i^2}{\pi_{io}} - 1\Big) = \sum_{i=1}^{s} \frac{F_i^2}{n\pi_{io}} - n. \tag{5.20}$$

Example 5.5 (Continued) The die is rolled 60 times. The frequencies with which the six faces of the die showed up are shown in Table 5.5. The question, once again, is: Is the die fair?

Table 5.5
Observations, Example 5.5

Face (C_i)	Frequency (F_i)	Relative frequency (R_i)
1	11	0.183
2	9	0.150
3	12	0.200
4	8	0.133
5	9	0.150
6	11	0.183
		1.000

Let us examine the consequences of the two types of error. A Type I error in this case is associated with rejecting a fair die; a Type II error means accepting a bad die. We may assume the consequences to be far more serious in the latter case, and we should want $1-\alpha$ small, meaning α must be large. For the sake of this illustration, let us suppose $\alpha = 0.99$. Since $s = 6$, we get from Appendix 4H $V_{0.99;5} = 0.554$. The V statistic is calculated as follows:

$$\begin{aligned} V &= \sum_{i=1}^{s} \frac{F_i^2}{n\pi_{io}} - n \\ &= \frac{1}{(60)(1/6)}[(11)^2 + (9)^2 + \cdots + (11)^2] - 60 \\ &= 1.20. \end{aligned}$$

Since $V > 0.554$, the hypothesis that the die is fair is rejected.

Related to the above is the test of the hypothesis that *the population distribution has a given mathematical form with certain (unspecified) parameter values.* An example would be the test of the hypothesis that the distribution of the measurement of an independent process is normal with *some* values of the parameters μ and σ. (If the parameter values as well as the form of the distribution are specified—for example, by the hypothesis that the population distribution is normal with parameter $\mu = 2.0$ and $\sigma = 0.3$—the appropriate test is the earlier one in this section.)

In advanced mathematical statistics texts, it is shown that, *when the random sample is large and with replacement, or large and from an independent process*, the decision rule for this test is given by (5.18), except that the critical value is $V_{\alpha;s-k-1}$, where k is the number of estimated parameters of the hypothesized distribution. In calculating the V statistic, π_{io} are probabilities determined under the assumption that the form of the distribution is that specified by the hypothesis, with appropriately estimated parameter values.

Example 5.6 Five hundred items produced by an independent manufacturing process were selected and the number of defects in each determined. The results are shown in columns (1), (2), and (3) of Table 5.6. Column (4) will be explained shortly.

The question is: Can it be supposed that the distribution of the number of defects per item in the process is Poisson?

If the distribution of the number of defects is Poisson with parameter m, the proportions of items with $x = 0, 1, 2, \ldots$ defects in the long run would be given by

$$p(x) = \frac{m^x e^{-x}}{x!}. \tag{5.21}$$

Table 5.6
Distribution of number of defects, Example 5.6

Number of defects (1)	Number of items, F_i (2)	Proportion of items, R_i (3)	Proportion under H_1, π_{io} (4)
0	463	0.926	0.9213
1	34	0.068	0.0755
2	2	0.004	0.0031
3	1	0.002	0.0001
	500	1.000	1.0000

The single parameter of the Poisson distribution, m, happens to be equal to the mean of the distribution (see Table 2.3 of Chapter 2). It makes sense, therefore, to estimate m by the sample mean—in this case, the average number of defects per item, $\bar{X}$:

$$\bar{X} = (0)(0.926) + (1)(0.068) + (2)(0.004) + (3)(0.002) = 0.082.$$

After substituting $m = 0.082$ in Equation (5.21), we can calculate $p(x)$ for $x = 0, 1, 2, \ldots$. A computer program was used to generate the figures shown in column (4) of Table 5.6. The last entry in column (4) is the sum of $p(x)$ for $x \geq 3$.

To test the hypotheses:

$$\begin{aligned} H_1 &: \text{ Process distribution of defects is Poisson,} \\ H_2 &: \text{ Process distribution of defects is not Poisson,} \end{aligned}$$

we first calculate the V statistic applying (5.19):

$$V = 500\Big[\frac{(0.9260 - 0.9213)^2}{0.9213} + \cdots + \frac{(0.0020 - 0.0001)^2}{0.0001} = 18.57.$$

Let us examine the possible consequences of the two types of error. A Type I error is associated with rejecting the hypothesis that the distribution is Poisson when in fact it is; a Type II error is that associated with accepting the Poisson hypothesis when it is false. Now, we did not explain why we wanted to test this hypothesis in the first place, but the reader may be willing to accept without elaboration that certain elegant and useful results in statistical quality control are applicable in the case where the process distribution is Poisson. Applying these results in the belief that H_1 is true when in fact it is not is probably the more serious error. For this illustration, let us suppose that the probability of a Type II error should not exceed 0.25, implying $\alpha = 0.75$.

One parameter (m) was estimated under H_1. Therefore, $k = 1$, $s - k - 1 = 4 - 1 - 1 = 2$, and $V_{\alpha;s-k-1} = V_{0.75;2} = 0.575$. Since $V > V_{\alpha;s-k-1}$, H_1 is rejected.

The first of the tests of this section, it should be noted, specifies H_1 *precisely.* Although it apppears appropriate for the example used to illustrate it (testing a die for fairness), it should not be applied in cases where it is unlikely that the population or process satisfies H_1 *exactly.* In such a case, there would be no point carrying out the test—it would either reject H_1, as expected, or would confirm that the sample size is not large enough to reject it.

The second test of this section, though less restrictive than the first, shares the same problem. For it may well be that the issue is not whether the process is exactly—say—Poisson (a feature which few real-world processes possess precisely), but whether it approximates the Poisson well enough for the purposes of the study.

5.8 TESTING THE INDEPENDENCE OF TWO ATTRIBUTES OR VARIABLES

Imagine the elements of a population classified into s categories, values, or intervals A_1, A_2, ..., A_s, according to one attribute or variable, and into t categories or intervals B_1, B_2, ..., B_t, according to a second. Let π_{ij} be the proportion of elements in the population which fall into categories A_i and B_j. Table 5.7 illustrates the notation.

Table 5.7
Population joint distribution

First	Second attribute			
attribute	$\cdots$	B_j	$\cdots$	Total
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	$\cdots$	π_{ij}	$\cdots$	$\pi_{i.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Total	$\cdots$	$\pi_{.j}$	$\cdots$	1.0

The marginal relative frequency of category A_i is denoted by $\pi_{i.}$, while that of B_j by $\pi_{.j}$.

Two attributes or variables are independent if all joint relative frequencies equal the product of the corresponding marginals, that is, if $\pi_{ij} = \pi_{i.}\pi_{.j}$ for all i and j. (The notation is different, but the definition of independence is identical to that in Chapters 1 and 2.)

Instead of a population of finite size, we may have an independent process, the elements of which are characterized by two measurements (for example, the length and width of rectangular metal plates produced by a stamping machine). By analogy with the one-measurement case, we shall say that a two-measurement process is independent if the joint distribution of measurements does not vary from element to element, and if one element's joint measurements are not related to those of any other element.

How, then, are we to determine whether or not the attributes or measurements are independent?

Suppose that a sample is drawn from such a population or process. Let R_{ij} be the proportion of the sample elements falling into categories A_i and B_j. Also, let $R_{i.}$ and $R_{.j}$ be the marginal relative frequencies of categories A_i and B_j respectively, as illustrated in Table 5.8.

Table 5.8
Sample joint distribution

First attribute	Second attribute $\cdots$	B_j	$\cdots$	Total
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	$\cdots$	R_{ij}	$\cdots$	$R_{i.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Total	$\cdots$	$R_{.j}$	$\cdots$	1.0

Before stating the decision rule for this case, let us illustrate the notation with an example.

Example 5.7 It has been argued that the main difference between younger and older drivers is the tendency of the former to have relatively more accidents and claims. However, again according to this argument, the amount of the claim is determined largely by the circumstances of the accident and should be unrelated to the age of the driver. On the other hand, if younger drivers tend to drive larger and more expensive cars faster, the severity of any accident in which they are involved will tend to be greater, and the claim amount should be related to age.

The most recent 500 claims received by an automobile insurance company were analyzed, and the joint relative frequency distribution of claim amount and age of the insured was obtained, as shown in Table 5.9.

For example, in 1.4% of the 500 selected files the insured was under 30 and the amount of the claim was over \$10,000.

The question is: Is the amount of the claim independent of the age of the insured?

Table 5.9
Distribution of age and claim amount,
Example 5.7

	Claim amount ($000)			
Age	Under 1	1 to 10	Over 10	Total
Under 30	0.362	0.016	0.014	0.392
30 to 50	0.318	0.024	0.012	0.354
Over 50	0.240	0.008	0.006	0.254
Total	0.920	0.048	0.032	1.000

The receipt of claims by the insurance company could be assumed to form an independent random process (why?) characterized by several measurements: the amount of the claim, the type of claim, the sex and age of the insured, etc.

The hypotheses to be examined are:

H_1 : Age and claim amount are independent
H_2 : They are not

Note that the hypotheses refer to the *process* from which the files are selected. A simple calculation will show that the definition of independence is not satisfied for the sample.

In the general case, if the attributes or measurements are independent, the sample joint relative frequencies, R_{ij}, will tend to be equal to the product of the corresponding marginals, $R_{i.}R_{.j}$, in which case the statistic

$$V = n \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(R_{ij} - R_{i.}R_{.j})^2}{R_{i.}R_{.j}} \tag{5.22}$$

will tend to be close to zero. If, on the other hand, the attributes are not independent, the R_{ij} will tend to deviate from the $R_{i.}R_{.j}$, and the value of V will tend to be large.

A reasonable decision rule, then, is to accept the hypothesis of independence when V is small, and to reject it when V is large. This decision rule can be written as

Accept H_1 if $V \leq c$,
Reject H_1 if $V > c$,

where c distinguishes small from large values of V.

For a large sample from an independent process or with replacement, the approximate decision rule for testing

$$\begin{aligned} H_1 &: \pi_{ij} = \pi_{i.}\pi_{.j}, \quad \text{for all } i \text{ and } j, \\ H_2 &: \pi_{ij} \neq \pi_{i.}\pi_{.j}, \quad \text{for some } i \text{ and } j, \end{aligned} \tag{5.23}$$

so that the probability of a Type I error equals α and that of a Type II error does not exceed $1-\alpha$, is to

$$\begin{aligned} &\text{Accept } H_1 \text{ if } V \leq V_{\alpha;(s-1)(t-1)}, \\ &\text{Reject } H_1 \text{ if } V > V_{\alpha;(s-1)(t-1)}, \end{aligned} \tag{5.24}$$

where V is given by (5.22), s is the number of rows and t the number of columns of Tables 5.7 and 5.8, and $V_{\alpha;(s-1)(t-1)}$ is obtained from Appendix 4H.

In mathematical statistics, it is shown that *if H_1 is true, the sample large, and either random and with replacement from a finite population, or selected from an independent process*, the probability distribution of V is approximately chi-square with parameter λ equal to $(s-1)(t-1)$. The probability of a Type I error will be equal to α, if c is set to $V_{\alpha;(s-1)(t-1)}$. This decision rule is summarized in the box.

Two comments should be made before illustrating this test. First, observe that the V statistic is closely related to the coefficient of association (P) of Chapter 1, which measures the strength of the relationship between two attributes. In fact, $V = n(q-1)P$, where q is the smaller of the number of rows and columns of Table 5.8.

Second, since joint relative frequencies are related to frequencies, $R_{ij} = F_{ij}/n$, V can also be written as

$$V = \sum_i \sum_j \frac{(nR_{ij} - nR_{i.}R_{.j})^2}{nR_{i.}R_{.j}} = \sum_i \sum_j \frac{(F_{ij} - E_{ij})^2}{E_{ij}}, \tag{5.25}$$

where, in the second term, $E_{ij} = nR_{i.}R_{.j}$.

Example 5.7 (Continued) The V statistic is calculated from the data in

Table 5.9, as follows:

$$V = 500[\underbrace{\frac{(0.362 - 0.920 \times 0.392)^2}{0.920 \times 0.392} + \cdots + \frac{(0.006 - 0.032 \times 0.254)^2}{0.032 \times 0.254}}_{9\ terms}]$$
$$= 2.938.$$

Now, a Type I error in this case is the rejection of the hypothesis of independence when it is in fact true. A Type II error is the conclusion that age and claim amount are unrelated when in fact they are not.

If we happen to advocate that age and claim amount are independent, and prefer to guard against a Type I error by making α small—say, $\alpha = 0.10$, we would find in Appendix 4H that $V_{0.10;4} = 7.779$, and would accept H_1.

If we happen to advocate the opposite theory and like to guard against a Type II error, we would prefer $1 - \alpha$ low—say, $1 - \alpha = 0.10$. This implies $\alpha = 0.90$, and $V_{0.90;4} = 1.064$. The decision then would be to reject the hypothesis of independence.

The two conclusions are based on the same evidence. The opposing interests lead to an opposite interpretation of this evidence.

It must be kept in mind that the hypothesis of independence is very strict. For H_1 in (5.23) to be true, *all* π_{ij} *must equal exactly* $\pi_{i.}\pi_{.j}$ (to the tenth, thousandth, ... decimal place). In the real world, it is very rare that two variables or attributes satisfy precisely this definition. In most situations, in other words, the test of this section serves no useful purpose because it is applied to a hypothesis that is known to be false; any failure to reject the hypothesis is simply a consequence of not having a large enough sample. In such cases, the question should not be whether or not the variables or attributes are related, but whether the relationship is strong enough for practical purposes. In the last example, for instance, an "acceptance" of the hypothesis that age and claim amount are independent is in practical terms "non-rejection." It should be interpreted to mean that the sample is not large enough to reject the hypothesis of independence. If the test rejects the hypothesis of independence, one should examine whether or not the relationship is too weak to matter.

5.9 COMPARING TWO OR MORE POPULATIONS

Suppose that a random sample is drawn from each of t populations. The elements of each sample are classified into s common categories, values, or intervals $C_1, C_2, \ldots, C_s$, according to the *same* attribute or variable. Let n_j be the size of the jth sample, F_{ij} the number, and R_{ij} the proportion

Table 5.10
Frequency distributions of t samples

Category, C_i	Sample 1	Sample 2	$\cdots$	Sample t	Total	Aggregate relative frequencies, $R_{i.}$
C_1	F_{11}	F_{12}	$\cdots$	F_{1t}	$F_{1.}$	$R_{1.} = F_{1.}/n$
C_2	F_{21}	F_{22}	$\cdots$	F_{2t}	$F_{2.}$	$R_{2.} = F_{2.}/n$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
C_s	F_{s1}	F_{s2}	$\cdots$	F_{st}	$F_{s.}$	$R_{s.} = F_{s.}/n$
Total	n_1	n_2	$\cdots$	n_t	n	1

of elements of the jth sample that fall into category C_i. The notation is illustrated in Table 5.10.

The problem is: Are the populations from which the samples were drawn identical with respect to the distribution of the variable or attribute?

Example 5.8 A consumer product testing organization investigated the durability of four brands of alkaline D batteries. The organization instructed each of its shoppers (widely scattered across the country) to purchase a few batteries at a large store. The batteries were then shipped to the central laboratory, where they were tested for durability under uniform conditions. In all, 100 Brand A, 120 Brand B, 80 Brand C, and 200 brand D batteries were tested, with the results shown in Table 5.11.

Table 5.11
Four brands of batteries compared

Life (hours)	Brand A	Brand B	Brand C	Brand D	Total	$R_{i.}$
Under 19 hr	11	10	8	21	50	0.100
19 to 20 hr	29	35	25	63	152	0.304
20 to 21 hr	42	50	30	77	199	0.398
Over 21 hr	18	25	17	39	99	0.198
Total	100	120	80	200	500	1.000

For example, of the 200 Brand D batteries tested, 21 lasted under 19 hours, 63 lasted between 19 and 20 hours, and so on. Note that the sample sizes are not (and need not be) the same. The question is: Are the batteries of the four brands identical in terms of durability? In different words: Are the population distributions of battery life the same for all brands?

The sample, it will be noted, is not random in the strict sense. There is obviously no list of batteries from which a random sample could be selected, and physical randomization is impossible. On the other hand, it could be argued that the organization's method produces an essentially random sample, in that the batteries actually selected had the same chance of appearing in the sample as any other batteries. This will be assumed here, but it is instructive for the reader to consider reasons why the sample should not be considered an essentially random one.

Returning to the general case, the problem is to test the hypothesis that the populations from which the t samples are drawn have identical distributions; that is, that the proportions of elements of each population which fall into categories $C_1, C_2, \ldots, C_s$ are the same—and equal to, say, $\pi_1, \pi_2, \ldots, \pi_s$.

If the populations are identical, the t separate samples may as well be combined into one, assumed drawn from the pooled populations. Three consequences may be expected to follow: (a) the aggregate relative frequencies of categories C_1, C_2, ..., C_s in the combined sample ($R_{1.} = F_{1.}/n$, $R_{2.} = F_{2.}/n$, ..., $R_{s.} = F_{s.}/n$) will tend to be close to the population relative frequencies π_1, π_2, ..., π_s of these categories; (b) the expected frequency of category i in sample j, $n_j\pi_i$, will tend to be close to $E_{ij} = n_j R_{i.}$, and this, in turn, will tend to be close to the observed sample frequency, F_{ij}; finally, (c) the value of the statistic

$$V = \sum_{i=1}^{s}\sum_{j=1}^{t} \frac{(F_{ij} - E_{ij})^2}{E_{ij}} \tag{5.26}$$

will tend to be small.

On the other hand, if the populations are not identical, the sample frequencies will tend to deviate from the E_{ij}, and V will tend to be large. It is reasonable, therefore, to accept H_1 when V is small, and to reject it when V is large.

In mathematical statistics, it is shown that when the populations are identical and the samples are large and with replacement, the probability distribution of the statistic V is approximately chi-square with parameter equal to $(s-1)(t-1)$. The decision rule summarized in the box that follows is based on this last result. In the box, π_{ij} denotes the proportion of elements of population j that belong to category C_i.

The test of (5.27) is identical to that of the previous section. It should be realized, however, that Tables 5.8 and 5.10 represent two different situations. Table 5.8 shows a *single* sample of size n drawn from a single

If each of t random samples with replacement or from independent processes is large, the approximate decision rule for testing

$$H_1 : \begin{cases} \pi_{11} = \pi_{12} = & \cdots & = \pi_{1t} = \pi_1 \\ \pi_{21} = \pi_{22} = & \cdots & = \pi_{2t} = \pi_2 \\ & \cdots & \\ \pi_{s1} = \pi_{s2} = & \cdots & \pi_{st} = \pi_s \end{cases} \tag{5.27}$$

against H_2 that at least one of the above equalities does not hold, so that the probability of a Type I error equals α and that of a Type II error does not exceed $1 - \alpha$, is to

$$\begin{aligned} \text{Accept } H_1 \text{ if } V \leq V_{\alpha;(s-1)(t-1)}, \\ \text{Reject } H_1 \text{ if } V > V_{\alpha;(s-1)(t-1)}. \end{aligned} \tag{5.28}$$

V is given by (5.26), and $V_{\alpha;(s-1)(t-1)}$ is tabulated in Appendix 4H.

population, classified jointly according to *two* attributes. Table 5.10, on the other hand, shows the results of t random samples, each drawn from a different population, but all commonly classified into s categories according to a *single* attribute.

Example 5.8 (Continued) If the four samples come from populations having identical distributions, the estimates of the common population relative frequencies of the four categories are shown in the last column of Table 5.11: $R_{1.} = 0.100$, $R_{2.} = 0.304$, $R_{3.} = 0.398$, and $R_{4.} = 0.198$. The estimated expected frequencies under H_1, $E_{ij} = n_j R_{i.}$, are shown in Table 5.12.

Table 5.12
Estimated expected frequencies, E_{ij}, Example 5.8

Category	Brand A	Brand B	Brand C	Brand D
Under 19 hr	10.0	12.00	8.00	20.0
19 to 20 hr	30.4	36.48	24.32	60.8
20 to 21 hr	39.8	47.76	31.84	79.6
Over 21 hr	19.8	23.76	15.84	39.6
Total	100.0	120.00	80.00	300.0

If the samples come from identical populations, the estimate of the proportion of batteries that last, say, between 20 and 21 hours is 0.398. Therefore, 39.8 of the 100 Brand A batteries, 47.76 of the 120 B batteries, 31.84 of the 80 C batteries, and 79.6 of the 200 D batteries can be expected to last between 20 and 21 hours.

We now calculate the V statistic (the F_{ij} come from Table 5.11, and the E_{ij} from Table 5.12):

$$\begin{aligned} V &= \sum_{i=1}^{4}\sum_{j=1}^{4}\frac{(F_{ij}-E_{ij})^2}{E_{ij}} \\ &= \underbrace{\frac{(11-10)^2}{10}+\frac{(29-30.4)^2}{30.4}+\cdots+\frac{(39-39.6)^2}{39.6}}_{16\ terms} \\ &= 1.447. \end{aligned}$$

Let us examine the consequences of the two types of error. If the distribution of battery life is identical for all four brands, and the consumer organization reaches the opposite conclusion, a Type I error is made. A Type II error, in this case, is the conclusion that the distributions are identical when in fact they are not. Which error is more serious? It is difficult to say, is it not? Let us assume that the two errors are thought to be equally serious, in which case α may be set at 0.50.

Since $s = 4$ and $t = 4$, $(s-1)(t-1) = 9$. For $\alpha = 0.50$, Appendix 4H gives $V_{0.50;9} = 8.343$.

The value of the test statistic is smaller than the critical value of the test, and we accept the hypothesis that the population distributions of the four brands are identical in terms of durability.

Once again note that H_1 is a very precise statement. To assert that the population distributions are identical means, strictly speaking, that there is absolutely no difference among the relative frequencies of any category. In this example, the acceptance of H_1 is best interpreted as non-rejection: the samples are not large enough to determine conclusively that the populations are not identical.

5.10 COMPARING TWO POPULATION MEANS

The test described in the last section is that of the hypothesis that the populations from which the samples are drawn have identical distributions. Occasionally, however, it may be useful to examine whether or not a certain characteristic *only* (such as the mean or the variance) of these population distributions is the same. Obviously, these are different questions. For example, if two distributions have equal means but different variances, their distributions are different. On the other hand, if two distributions are identical, their means, variances, and all other characteristics are the same.

We shall describe here a test of the hypothesis that *two population means are equal.* The first population consists of N_1, and the second of N_2 elements. Let μ_1 and μ_2 be the means of a certain variable X in the two populations. At issue is whether or not $\mu_1 = \mu_2$ (alternatively, whether or not $\mu_1 - \mu_2 = 0$).

A random sample is drawn from each of the two populations. Let n_i, $\bar{X}_i$, and S_i^2 be the size of the sample, and the mean and variance of variable X in the sample drawn from population i ($i = 1, 2$).

A statistic on which the test can be based is the difference between the two sample means, $\bar{X}_1 - \bar{X}_2$. If the two population means are equal, the difference $(\bar{X}_1 - \bar{X}_2)$ will tend to be close to 0; if not, $(\bar{X}_1 - \bar{X}_2)$ will tend to deviate from 0. We shall want to accept H_1 when $(\bar{X}_1 - \bar{X}_2)$ is close to 0, and to reject it when $(\bar{X}_1 - \bar{X}_2)$ is not close, in either direction, to 0. This decision rule can be written as

$$\begin{array}{ll} \text{Accept } H_1 \text{ if} & |\bar{X}_1 - \bar{X}_2| \leq c, \\ \text{Reject } H_1 \text{ if} & |\bar{X}_1 - \bar{X}_2| > c, \end{array}$$

where c, the "critical value" of the test, is a number distinguishing "close to 0" from "not close enough to 0" values. As usual, the problem is to determine c so that the probability of a Type I error does not exceed α, while that of a Type II does not exceed $1 - \alpha$. It can be shown that, when the population and sample sizes are large, the approximate decision rule is that described in the box that follows.

Example 5.9 The consumer product testing organization of Example 5.8 also investigated the durability of two major brands of heavy-duty zinc chloride batteries. A total of 100 Brand A and 120 Brand B batteries were purchased throughout the country and tested under uniform conditions with the following results.

	Brand A	Brand B
Sample size	$n_1 = 100$	$n_2 = 120$
Average life (hours)	$\bar{X}_1 = 14.25$	$\bar{X}_2 = 10.63$
Standard deviation	$S_1 = 1.92$	$S_2 = 1.81$

At issue is whether or not the *average life* of all Brand A batteries is identical to that of all Brand B batteries.

Assuming, for the reasons stated in Example 5.8, that the samples can be considered essentially random, we calculate first:

$$|\bar{X}_1 - \bar{X}_2| = 3.62.$$

For random samples without replacement, and large n_1, N_1, n_2, and N_2, the approximate decision rule for testing

$$\begin{aligned} H_1 &: \mu_1 = \mu_2, \\ H_2 &: \mu_1 \neq \mu_2, \end{aligned} \tag{5.29}$$

so that $Pr(\text{Type I error}) = \alpha$ and $Pr(\text{Type II error}) \leq 1 - \alpha$, is to

$$\begin{aligned} \text{Accept } H_1 \text{ if } \quad |\bar{X}_1 - \bar{X}_2| \leq c, \\ \text{Reject } H_1 \text{ if } \quad |\bar{X}_1 - \bar{X}_2| > c, \end{aligned} \tag{5.30}$$

where

$$c = U_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}\frac{N_1 - n_1}{N_1 - 1} + \frac{S_2^2}{n_2}\frac{N_2 - n_2}{N_2 - 1}}. \tag{5.31}$$

Selected values of $U_{\alpha/2}$ are listed in Table 5.1.

If, as we assumed in Example 5.8, the two types of error are thought to be equally serious and $\alpha = 0.50$, then $U_{\alpha/2} = 0.674$. Next, we calculate c. Since the N_i are very large, we may set $(N_i - n_i)/(N_i - 1) \approx 1$, in which case

$$c = 0.674\sqrt{\frac{(1.92)^2}{100} + \frac{(1.81)^2}{120}} = 0.674\sqrt{0.064} = (0.674)(0.253) = 0.171.$$

Since $3.62 > 0.171$, we reject the hypothesis that the average life of the two brands is the same.

Once again, the strictness of H_1 should be kept in mind: for H_1 to be true, μ_1 must equal μ_2 *precisely*. There is no point in applying this test in situations where this equality cannot possibly be true.

It is sometimes useful to construct a confidence interval for the difference between two population means. In advanced statistics, it is shown that *for large n_1, N_1, n_2, and N_2, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by*

$$(\bar{X}_1 - \bar{X}_2) \pm c, \tag{5.32}$$

with c given by Equation (5.31) above. The meaning is similar to that of all confidence intervals: the probability is $1-\alpha$ that the interval (5.32) contains the true difference $\mu_1 - \mu_2$.

Example 5.9 (Continued) A 95% confidence interval for the difference in the average life of the two brands of batteries is

$$(3.62) \pm (1.96)(0.253),$$

or from about 3.12 to 4.12 hours.

5.11 A NON-PARAMETRIC TEST

Let us consider once again the problem of comparing the durability of two brands of batteries. Let us suppose that samples of five Brand A and 6 Brand B batteries were obtained in the manner described in Example 5.8 and tested, with the results shown in columns (1) and (2) of Table 5.13. (The samples are deliberately small to illustrate simply the calculations that follow.)

Table 5.13
Samples of two brands of batteries

Life (hours)		Ranks	
Brand A (1)	Brand B (2)	Brand A (3)	Brand B (4)
	8.9		1
	9.9		2
10.1		3	
	10.2		4
	10.5		5
12.4		6	
	12.7		7
	13.6		8
14.2		9	
16.8		10	
17.9		11	
n_1=5	$n_2 = 6$	$R_1 = 39$	$R_2 = 27$

Let us combine the two samples, arrange the observations in increasing order of magnitude, and assign to them the ranks shown in columns (3) and (4) of Table 5.13.

If the two populations from which the samples were drawn are identical (that is, if the distributions of battery life are the same for the two brands), the ranks 1 to 11 would tend to be evenly distributed between the two brands; neither column (3) nor column (4) should show a concentration of low or high ranks. In fact, if the populations are identical, the sum of the ranks in column (3) should behave like the sum of five integers selected at random and without replacement from among the first 11 (and that of column (4) like the sum of 6 such integers).

In general, assume the sample from the first population consists of n_1 and that from the second of n_2 elements. Let R_1 denote the sum of ranks assigned to the first sample. If the two populations are identical, it can be shown that the expected value and variance of R_1 in random samples of size n_1 and n_2 *with replacement* are:

$$E(R_1) = \frac{n_1(n_1+n_2+1)}{2},$$
$$Var(R_1) = \frac{n_1 n_2(n_1+n_2+1)}{12}.$$

It can also be shown, again if the two populations are identical, that the probability distribution of R_1 is approximately normal for large n_1 and n_2. (It does not matter which population and sample is called "the first" and which "the second." R_1 could refer to either, but n_1 should be the corresponding sample size.)

If the populations are identical, the observed R_1 should tend to be close to $E(R_1)$; if not, R_1 should tend to deviate from $E(R_1)$. Therefore, we should want to reject the hypothesis that the two populations are identical when R_1 deviates substantially from $E(R_1)$. The terms "close" and "substantially" are given precise meaning in the decision rule shown in the box that follows.

To illustrate this test, let us return to our example and assume—as we did in Example 5.8—that $\alpha = 0.50$. Type I and II errors are assumed equally serious, and $U_{\alpha/2} = 0.674$. Calculate

$$E(R_1) = \frac{(5)(5+6+1)}{2} = 30, \quad Var(R_1) = \frac{(5)(6)(5+6+1)}{12} = 30.$$

$E(R_1) \pm U_{\alpha/2}\sqrt{Var(R_1)}$ is $(30) \pm (0.674)\sqrt{30}$, or from about 23.61 to 33.69. Since the observed $R_1 = 39$ lies outside this interval, H_1 is rejected.

This decision rule is known as the *Mann-Whitney* or *Wilcoxon (MWW) test*. Three points are worth noting.

First, the MWW test is an alternative to the chi-square test of Section 5.9. As we noted in Section 5.5, there may be more than one test of given hypotheses. This raises the question of how to determine which of the alternative tests is better—a question that has received considerable attention in the statistical literature, but which we shall not pursue here.

If sampling is with replacement or from an independent process, and the sample sizes n_1 and n_2 are large, the approximate decision rule for testing

$$H_1: \text{ Two populations are identical}$$
$$H_2: \text{ Two populations are not identical}$$

so that the probability of a Type I error equals α and that of a Type II error does not exceed $1-\alpha$, is to

$$\begin{aligned} &\text{Accept } H_1 \text{ if } E(R_1) - U_{\alpha/2}\sqrt{Var(R_1)} \le R_1 \\ &\qquad\qquad \le E(R_1) + U_{\alpha/2}\sqrt{Var(R_1)}, \\ &\text{Reject } H_1 \text{ if otherwise.} \end{aligned}$$

Selected values of $U_{\alpha/2}$ are given in Table 5.1.

The second point is that there is a MWW test applicable for *any*—not only large—n_1 and n_2; the critical values for this test are obtained from special tables, but the test itself is based on R_1. The noteworthy feature of this test is that it does not require the two population distributions to have a particular form. As we remarked earlier, decision rules applicable to samples of any size usually make this requirement; they are sometimes called *parametric* tests, in the sense that the hypotheses deal with parameters of population distributions of a given type. By contrast, the MWW is one of many *non-parametric* tests.

The final point worth noting is that H_1, once again, is a precise statement. What is being tested is the hypothesis that the two population distributions are *identical.* This non-parametric test, therefore, shares the main shortcoming of the chi-square test of Section 5.9. Since it is very rare that two population distributions in the business world are exactly alike, care should be taken that rejection of H_1 be followed by an appraisal of the magnitude of the differences between the two populations.

PROBLEMS

5.1 A random sample of size $n = 200$ was drawn without replacement from a population of size $N = 1,000$. The sample mean of a variable is $\bar{X} = 150$, and the sample variance is $S^2 = 280$.

(a) Calculate the 99%, 95%, 90%, and 50% symmetric confidence intervals for the population mean of the variable, μ. Briefly interpret these intervals.

(b) Calculate the 99%, 95%, 90%, and 50% symmetric confidence for the total value of the variable in the population. Briefly interpret these intervals.

(c) Do (a) and (b) under the assumption that the sample is with replacement.

5.2 A random sample of size $n = 200$ was drawn without replacement from a population of size $N = 1,000$. The proportion of elements in the sample falling into a certain category is $R = 0.43$.

(a) Calculate the 99%, 95%, 90%, and 50% symmetric confidence intervals for the proportion of elements in the population that fall into this category. Briefly interpret these intervals.

(b) Calculate the 99%, 95%, 90%, and 50% symmetric confidence intervals for the number of elements in the population that belong to this category. Briefly interpret these intervals.

(c) Do (a) and (b) under the assumption that the sample is with replacement.

5.3 In the manner described in Section 5.3, construct (a) a two-sided symmetric, (b) a two-sided asymmetric, and (c) two one-sided 90% confidence intervals for (i) the population mean of a variable, and (ii) a population proportion. Briefly discuss the differences among these intervals. Which type of interval is preferable?

5.4 Following complaints that parking meters were malfunctioning, the Department of Consumer Affairs selected at random and without replacement 10% of the 1,850 parking meters in a metropolitan area. Of the meters tested, 105 gave the correct reading, 75 gave more time that was paid for, and 5 gave less time than was paid for. "Parking meters," the Department's news release concluded, "may be one of the few bargains left."

What is your estimate of the proportion of all parking meters giving the correct time? Construct an interval estimate of this proportion; the interval estimate should contain the true proportion with probability 95%. Do you need any additional information in order to determine if parking meters are indeed a bargain? If so, what?

5.5 The management of a supermarket is concerned about the average waiting time of its customers at the checkout counters on Saturdays. Fifty customers "were randomly selected" one Saturday, and the time they spent waiting in line before being served was recorded. Their average waiting time was 5.2 minutes and the standard deviation of waiting times was 1.7 minutes.

(a) Assuming that the selected customers constitute a random sample with replacement from the population of all customers, test the hypothesis that the mean waiting time of *all* Saturday customers is less than or equal to 5 minutes, against the alternative hypothesis that it is greater than 5 minutes. Assume that the probability of a Type I error should not exceed 10%.

(b) Explain the exact meaning of Type I and II errors in this case, and their likely consequences. Which error is the more serious in this case?

(c) Redo the test in (a) under the assumption that the probability of a Type II error should not exceed 10%.

(d) How would you select a random sample of Saturday customers? Would *any* n customers, no matter how selected, constitute a sample to which the test in (a) could be applied?

5.6 The manager of a department store wished to estimate the proportion of time that the sales clerks are idle. She divided the week into three periods of about equal business volume (weekdays to 5 p.m., weekdays after 5 p.m., Saturdays). Over a period of one month, a number of checks were made in each period at times selected at random. A particular clerk was checked a total of 70 times during one of the three periods and in 12 of these times he was found to be idle.

(a) Assuming that the observations form a random sample from an infinite population of potential timings, test the hypothesis that the true proportion of time that the clerk is idle is less than or equal to 15%, against the alternative hypothesis that it is greater than 15%. The probability of a Type I error should not exceed 5%.

(b) Explain the exact meaning of Type I and II errors in this case, and their likely consequences. Which error is the more serious in this case?

(c) Redo the test in (a) under the assumption that the probability of a Type II error should not exceed 5%.

(d) How would you select the observation times so that the test in (a) can be applied?

5.7 The Plastics Packaging Division (PPD) manufactures plastic containers, such as cold-drink cups, yogurt cups, and creamers, by a high-speed extrusion and vacuum thermo-forming process. One of these products is an 8-oz yogurt container. The quality control department has established through tests the minimum wall thickness required to ensure that the container does not crack when filled. This minimum thickness is 11 mills (1 mill = 0.001 inches). A container with a wall thickness of less than 11 mills is considered defective, and one with wall thickness greater than or equal to 11 mills is considered good. These containers are produced at the rate of about 100,000 per hour, and are packaged for shipment to the customer in lots of about 40,000. Because of random variation in the quality of the raw material and the manufacturing process, it is impossible to ensure that *all* containers in a lot will be good. The Division and the customer agree that a lot is acceptable if no more than 4% of the containers are defective.

PPD's current quality control procedure is to take a random sample of 30 containers from each lot, and to reject the lot if the *average* wall thickness of the containers in the sample is less than 11 mills.

Comment on the current quality control procedure, and compare it critically with other possible procedures.

5.8 Cigarettes are manufactured by high-speed machines, some of which produce at a rate in excess of 5,000 per minute. These machines blend three main ingredients (tobacco lamina, tobacco stem, and synthetic or reprocessed tobacco), measure the quantity of blend that goes into each cigarette, and roll and wrap the cigarette with the proper paper.

A certain amount of variation in the finished weight of the cigarette is inevitable and tolerated, but the weight should be neither too low (because customers will find it unsatisfactory) nor too high (since this would tend to increase raw material costs). The ideal weight and tolerance for a particular type of cigarette is $1{,}000 \pm 100$ mg per cigarette. In other words, an acceptable cigarette should have a weight between 900 and 1,100 mg.

Brink Tobacco Company has purchased several high-speed machines. When a machine is functioning properly ("under control"), no less than 96% of the cigarettes produced are within the specification limits.

Occasionally, the machine goes "out of control." As a result, the proportion of cigarettes within the specification limits declines, while that of light and heavy cigarettes increases. The problem for Brink Tobacco is how to detect whether or not a machine is under control, and to be able to do so frequently, quickly, and economically. Obviously, it would be impossible to weigh every cigarette produced; whatever inspection system is used must be based on a sample.

It is proposed that a sample of 10 cigarettes be taken randomly from the machine's output at the stroke of every minute. Each cigarette in the sample will be weighed. Depending on weights observed, one of two decisions will be made: allow the machine to run, or shut it down in order to repair and adjust it.

(a) You are asked to formulate a plan such that the probability of shutting down a machine when it is under control does not exceed 20%. Without doing any calculations, describe how such a plan could be determined.

(b) Determine the plan in (a).

(c) Same as (b), but the sample size is large—say 200.

5.9 In order for a set of numbers to qualify as random numbers, it is necessary that the ten digits 0 to 9 appear with equal relative frequency (1/10) in the long run. Of course, the actual relative frequencies of a finite number of such random numbers will not always equal the theoretical relative frequencies.

One hundred numbers produced by a computer program yielded the frequency distribution shown in Table 5.14.

Table 5.14
Data, Problem 5.9

Number	Frequency
0	8
1	11
2	9
3	10
4	10
5	12
6	9
7	11
8	12
9	8
	100

(a) Assuming that the above numbers can be treated as a random sample from an infinite population of numbers that could be generated by this program, test the equal frequency hypothesis. The probability of a Type I error should not exceed 1%.

(b) Which other requirement must these numbers satisfy to qualify as random numbers? How could this be tested? Describe only, do not calculate.

(c) Determine the exact meaning of Type I and II errors in this case, and their likely consequences. Which error is more serious?

(d) Redo the test in (a) under the condition that the probability of a Type II error should not exceed 1%.

(e) Would *any* n numbers produced by the program qualify as a sample to which the tests in (a) or (d) apply?

5.10 A study was made of the time that elapsed between 150 successive telephone calls to a certain exchange. The results were as shown in Table 5.15.

(a) Assuming that the observations constitute a random sample from an infinite population of calls, test the hypothesis that the distribution of time between calls is exponential. The probability of a Type I error should not exceed 5%.

Table 5.15
Data, Problem 5.10

Time from previous call (minutes)	Number of calls
0.0 to 0.5	93
0.5 to 1.0	36
1.0 to 1.5	12
1.5 to 2.0	6
2.0 to 2.5	3
Total	150

Table 5.16
Wage distribution, Problem 5.11

Wage interval (dollars)	Number of workers
120 to 125	2
125 to 130	7
130 to 135	10
135 to 140	15
140 to 145	20
145 to 150	25
150 to 155	19
155 to 160	17
160 to 165	11
165 to 170	8
170 to 175	3
175 to 180	2
180 to 185	1
	140

Hints: Estimate the parameter λ of this distribution by the inverse of the average observed time between calls, using the midpoints of the time intervals above; that is, set $\lambda = 1/\bar{X}$ (why?). The probability that a variable X, having an exponential distribution with parameter λ, will be in the interval from a to b $(a < b)$ can be shown to be equal to

$$Pr(a \leq X \leq b) = e^{-\lambda a} - e^{-\lambda b}.$$

(b) Determine the exact meaning of Type I and II errors in this case, and their likely consequences. Which error is more serious?

(c) Redo (a) under the condition that the probability of a Type II error should not exceed 5%.

(d) Can the observations be assumed to form a sample to which the tests above may be applied?

5.11 The distribution of daily wages in a random sample without replacement of 140 workers in a large factory employing about 15,000 workers is shown in Table 5.16.

(a) Assuming that the sample can be treated as essentially one with replacement (because of the large population size), test the hypothesis that the distribution of wages in the factory is normal. The probability of a Type I error should not exceed 10%. *Hint:* Estimate the parameters μ and σ of the normal distribution by the sample mean and standard deviation of wages calculated using the midpoints of the intervals above.

(b) Determine the exact meaning of Type I and II errors in this case, and their likely consequences. Which error is more serious?

(c) Redo the test in (a) under the condition that the probability of a Type II error should not exceed 10%.

5.12 A study was made of the feasibility of constructing short-term storage facilities for liquid products (animal, vegetable, and marine oils, chemicals, etc., but excluding petroleum products) exported from or imported to a Great Lakes port. Table 5.17 shows the frequency distribution of the time between successive arrivals of all liquid-carrying vessels at the port in the most recent summer season.

Table 5.17
Great Lakes study, Problem 5.12

Interarrival time (days)	Number of vessels
0 to 5	23
5 to 11	10
11 to 15	6
Over 15	1*
	40

*24 days

Can it be assumed that the distribution of interarrival times at the port is exponential? Note the *Hints* of Problem 5.10, but otherwise answer this question as you consider appropriate.

5.13 One hundred thirty-eight students were enrolled in Administration 531 and 532 (Quantitative Methods, I and II) at an MBA program. Of these students, 79 received an A or B grade in *both* 531 and 532. The actual distribution of grades was as follows:

	ADM532		
ADM531	A	B	Total
A	18	16	34
B	31	14	45
Total	49	30	79

For example, 31 students had a B in ADM531 and an A in ADM532, etc.

As a person knowledgeable in statistical methods, you are invited to comment on these numbers.

5.14 Among the questions included in a survey of apartment tenants were the following:

A: Do you prefer living in an apartment? Yes___ No___
B: Do you have children? Yes___ No___
The following table summarizes the responses of a random sample of 100 tenants.

	B		
A	Yes	No	Total
Yes	15	35	50
No	25	25	50
Total	40	60	100

(a) Assuming that the population of tenants is very large so that the sample is essentially one with replacement, test the hypothesis that a tenant's preference is independent of the presence of children. The probability of a Type I error should not exceed 10%.

(b) Determine the meaning and likely consequences of the two types of error in this case. Which error is the more serious?

(c) Redo the test in (a) under the assumption that the two errors are equally serious.

5.15 A sample of 280 households in a large metropolitan area was selected in order to investigate the usage of XL White, a brand of laundry bleach. The results of the study are in part as shown in Table 5.18.

Table 5.18
Results of bleach study, Problem 5.15

Household income	Non-users	Light users	Heavy users	Total
Under 15,000	28	16	10	54
15,001 to 20,000	32	15	8	55
20,001 to 30,000	27	14	12	53
30,001 to 40,000	31	17	11	59
Over 40,000	32	18	9	59
Total	150	80	50	280
$V = 1.826$				
Family size	**Non-users**	**Light users**	**Heavy users**	**Total**
Under 3	60	26	10	96
3 to 4	52	33	15	100
Over 4	38	21	25	84
Total	150	80	50	280
$V = 13.799$				

(a) Test the two hypotheses that usage is independent of (i) income, and (ii) family size. The probability of a Type I error should not exceed 10%.

(b) Examine the meaning and likely consequences of the two types of error in this case. Which error is the more serious? Redo the tests under the condition that the probability of a Type II error should not exceed 10%.

(c) How would you select a simple random of households? Would *any* sample do for the purpose of applying the above tests?

(d) What are the implications of the test results?

5.16 A study was undertaken to determine the factors influencing handling time for metal plates used in a punch press. The weight of the metal piece was thought to be one of the determining factors. Accordingly, the weight category (light, medium, heavy) and the handling time (classified into short and long) were recorded for a random sample of 15 metal plates with the results shown in Table 5.19.

Table 5.19
Weight and handling time of metal plates, Problem 5.16

Plate No.:	Weight	Handling time
1	Medium	Long
2	Light	Short
3	Medium	Short
4	Heavy	Short
5	Light	Short
6	Heavy	Long
7	Medium	Short
8	Heavy	Long
9	Light	Short
10	Medium	Long
11	Light	Long
12	Medium	Short
13	Heavy	Long
14	Medium	Long
15	Heavy	Long

Does the weight of a metal piece influence the handling time? If yes, what is the nature of the relationship? In answering these questions treat the sample as if it were large. Why is this assumption necessary? Explain and justify any other assumptions you are forced to make.

5.17 A product testing laboratory was asked to evaluate the durability of four brands of tires. The durability tests were made under normal city driving conditions with the assistance of a firm operating a fleet of taxicabs. In all, 140 taxis were employed and each was fitted with four new tires, one from each brand. These tires were randomly selected from retail outlets. Each day, the tires were rotated in a predetermined manner to ensure uniform exposure to wear, unrelated to their original location on the car. Also daily, the tires were inspected to determine if they had reached the end of their useful life. The test results are summarized in Table 5.20.

(a) Assuming that the observations can be considered random samples from infinitely large populations, do the four samples of tires come from populations having identical distributions of life? The probability of a Type I error should not exceed 5%.

(b) Interpret the meaning of the two types of error in this case, and their likely consequences. Which error is more serious?

(c) Redo the test in (a) under the condition that the probability of a Type II error should not exceed 5%.

Table 5.20
Tire test results, Problem 5.17

Life (000 miles)	Brands A	B	C	D
Under 30	25	30	25	26
30 to 31	40	45	45	43
31 to 32	60	55	50	55
Over 32	15	10	20	16
Total	140	140	140	140

(d) Comment on the method used by the laboratory to measure the durability of tires. Does the method produce samples to which the tests in (a) and (c) may be applied?

5.18 In a sample of 200 beer consumers in city A, the average monthly beer consumption was 26.5 oz, and the sample standard deviation was 2.14. In a sample of 150 beer consumers in city B, the average monthly beer consumption was 29.8 oz, and the sample standard deviation was 3.87.

(a) Assuming that the samples are random and without replacement from very large populations, test the hypothesis that the average beer consumption of all beer drinkers in city A is the same as that in city B. The probability of a Type I error should not exceed 20%.

(b) Interpret the meaning and the likely consequences of the two types of error in this case. Which error is the more serious?

(c) Redo the test in (a) assuming the probability of a Type II error should not exceed 20%.

(d) How would you select a random sample of beer consumers in a city?

5.19 Surveys of potential buyers can provide useful information about the likelihood of success of *planned* (that is, as yet not manufactured or offered) products and services. In the case described here, the product in question was an electric car which, at the time the survey was conducted, was still at the prototype stage.*

Data were collected through personal interviews with 1,229 randomly selected shoppers at designated shopping centers in three cities. The respondents were given a detailed description of the planned electric car, which included its price, passenger and luggage capacity, size, speed, cost of operation, and safety features. The respondents were asked to rate their intention to buy the electric car when it became available on an 11-point scale, ranging from 0 (absolutely no chance of buying) to 10 (almost certain of buying). They were also asked to rate the importance they attributed to each of a number of factors influencing their choice of car on a 10-point scale, ranging from 0 (low) to 9 (high importance). The results of the study are shown in part in Tables 5.21, 5.22, and 5.23.

In Table 5.23, the importance groups in column (2) are formed as follows: I = low importance (0 to 3 on the 9-point scale); II = medium importance (4 to 6); and III = high importance (7 to 9). The intention groups are as follows: None = 0 or 1, Low = 2 to 4, Medium = 5 to 7, and High = 8 to 10 on the 11-point scale.

* This case is based on G. M. Naidu, G. Tesar, and G. Udell, "Determinants of buying intentions of the electric car," *1973 Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, pp. 515-20.

Table 5.21
Distribution of buying intention,
Problem 5.19

Rating	Frequency of response City A	City B	City C
0	94	53	33
1	35	12	16
2	32	12	31
3	43	12	38
4	28	16	37
5	101	65	63
6	37	30	29
7	58	36	35
8	53	32	26
9	23	31	12
10	52	33	21
Total	556	332	341
Mean	4.72	5.30	4.69
Std. dev.	3.22	3.21	2.78

Table 5.22
Mean and standard deviation
of ratings for factors, Problem 5.19

Factor	Ratings Mean	Std. deviation
Cost of operation	7.00	2.30
Ease of maintenance	6.24	2.49
Cost of maintenance	6.81	2.36
Luggage capacity	4.56	2.51
Passenger capacity	4.92	2.41
Size	5.18	2.58
Mileage	6.85	2.34
Price	6.91	2.33
Speed	4.34	2.43
Acceleration	4.69	2.58
Safety	7.24	2.36
Pollution	6.28	2.80

Column (7) shows the mean intention-to-buy rating for each importance group. Column (8) is the V statistic for testing the independence of factor and intention to buy.

Assume you are the marketing manager of the company intending to produce this electric car. Interpret these findings.

5.20 An issue of some importance in marketing is the extent to which consumers are conscious of the price of an article at the time of its purchase. One theory is that consumers watch prices and carefully adjust their purchases from different

Table 5.23
Association between importance of factor and intention to buy, Problem 5.19

Factor (1)	Imp. group (2)	None (3)	Intention group Low (4)	Medium (5)	High (6)	Mean intention (7)	V (8)
Cost	I	43	18	36	25	3.73	
of	II	49	83	92	50	4.18	67.19
operation	III	87	210	326	208	5.14	
Ease	I	47	40	78	38	4.18	
of	II	47	114	120	65	4.45	32.82
maintenance	III	85	157	256	180	5.13	
Cost	I	39	20	46	30	4.12	
of	II	49	85	101	56	4.29	37.12
maintenance	III	91	206	307	197	4.98	
	I	62	91	168	104	4.91	
Luggage	II	51	148	186	114	4.96	26.60
capacity	III	66	72	100	65	4.33	
	I	54	73	130	87	4.92	
Passenger	II	56	154	213	127	5.02	23.28
capacity	III	69	84	111	69	4.27	
	I	50	29	36	17	3.03	
Mileage	II	41	86	102	49	4.41	79.70
	III	88	196	316	217	5.19	
	I	35	30	37	20	3.59	
Price	II	49	96	95	59	4.40	29.99
	III	95	185	322	204	5.11	
	I	78	93	148	133	4.93	
Speed	II	59	144	218	103	4.68	31.76
	III	42	74	88	47	4.41	
	I	53	60	72	34	3.80	
Pollution	II	36	98	115	50	4.60	47.36
	III	90	153	267	199	5.18	

outlets so as always to minimize the total cost of the articles they buy. Opponents of this theory argue that consumers cannot remember accurately the prices of the hundreds of commodities they normally buy; they are often concerned more with whether or not they can afford to buy an article rather than with its exact price.

To investigate this issue, 640 housewives were approached at randomly selected addresses in a city.* "Housewife" was interpreted broadly to mean the person, male or female, responsible for current purchases of provisions for the household, but will be referred to as a "she" in this case. The questionnaire contained questions about recent purchases of fifteen selected commodities and about certain aspects of the household. For each commodity, the housewife was

* Adapted from A. Gabor and C. W. J. Granger, "On the price consciousness of consumers," *Applied Statistics*, Vol. 10, No. 3, pp. 170-88.

asked when she bought it last. The interviewers were instructed not to ask for further information if the last purchase was more than a week ago. However, if the housewife had purchased the commodity within the last week, she was asked to state the brand or type of the commodity, whether or not she recalled the price paid for it, and, if so, how much she paid. The interviewers were asked to supply their personal estimate of the social group to which the housewife belonged. Five social groups were distinguished: A (the well-to-do), B (the professional middle class), C (the lower middle class), D (the working class), and E (the poor). The recognizable characteristics of these groups were described in detail in the written instructions to the interviewers.

Of the fifteen commodities listed in the questionnaire, eight were sold at such a variety of prices that it was not practicable to check the answers. The remaining seven, however, could be checked and the price named by the housewife could be compared with the list price obtained from industry sources. If there was any departure from the list price, the answer was classified as incorrect. The Don't Know category includes both those who, for some reason, refused to name the price of the purchase, and those who admitted that the price named was more or less a guess. Altogether, 422 housewives reported 1,888 recent purchases of one or more of the seven commodities the price of which could be checked. The results by commodity are shown in Table 5.24.

Table 5.24
Results by commodity, Problem 5.20

Stated price was:	Commodity							
	Tea	Coffee	Sugar	Jam	Margarine	Flour	Cereal	All
Correct	283	109	266	117	116	100	85	1,076
Wrong	56	35	53	46	110	82	98	480
Don't know	18	16	78	34	26	99	61	332
All purchases	357	160	397	197	252	281	244	1,888

The relationship between the percentage of correct answers and social group is shown in Table 5.25.

Table 5.25
Results by social group, Problem 5.20

Social group	Number of housewives	Number of purchases	Number of prices named correctly	Percentage correct
A	9	42	19	45.2
B	28	116	54	46.6
C	118	544	309	56.8
D	229	1,052	616	58.6
E	38	134	78	58.2
All groups	422	1,888	1,076	57.0

The percentage correct is the ratio of the number of correct prices named to the number of items bought.

What are the implications of this study for the price-awareness theory?