# Statistics / Econometrics: Hypothesis Testing

**Course : Introductory Econometrics / Statistics : HC43**

B.A. Hons Economics, Semester IV / III

Delhi University

Course Instructor:

Siddharth Rathore

Assistant Professor

Economics Department, Gargi College

# APPENDIX D

# STATISTICAL INFERENCE: ESTIMATION AND HYPOTHESIS TESTING

Equipped with the knowledge of probability; random variables; probability distributions; and characteristics of probability distributions, such as expected value, variance, covariance, correlation, and conditional expectation, in this appendix we are now ready to undertake the important task of **statistical inference.** Broadly speaking, statistical inference is concerned with drawing conclusions about the nature of some population (e.g., the normal) on the basis of a random sample that has supposedly been drawn from that population. Thus, if we believe that a particular sample has come from a normal population and we compute the sample mean and sample variance from that sample, we may want to know what the true (population) mean is and what the variance of that population may be.

## D.1 THE MEANING OF STATISTICAL INFERENCE[1]

As noted previously, the concepts of *population* and *sample* are extremely important in statistics. *Population*, as defined in Appendix A, is the totality of all possible outcomes of a phenomenon of interest (e.g., the population of New York City). A *sample* is a subset of a population (e.g., the people living in Manhattan, which is one of the five boroughs of the city). Statistical inference, loosely speaking, is the study of the relationship between a population and a sample drawn from that population. To understand what this means, let us consider a concrete example.

[1]Broadly speaking, there are two approaches to statistical inference, Bayesian and classical. The classical approach, as propounded by statisticians Neyman and Pearson, is generally the approach that a beginning student in statistics first encounters. Although there are basic philosophical differences in the two approaches, there may not be gross differences in the inferences that result.

**TABLE D-1** PRICE TO EARNINGS (P/E) RATIOS OF 28 COMPANIES ON THE NEW YORK STOCK EXCHANGE (NYSE)

| Company | P/E | Company | P/E |
|---|---|---|---|
| AA | 27.96 | INTC | 36.02 |
| AXP | 22.90 | IBM | 22.94 |
| T | 8.30 | JPM | 12.10 |
| BA | 49.78 | JNJ | 22.43 |
| CAT | 24.68 | MCD | 22.13 |
| C | 14.55 | MRK | 16.48 |
| KO | 28.22 | MSFT | 33.75 |
| DD | 28.21 | MMM | 26.05 |
| EK | 34.71 | MO | 12.21 |
| XOM | 12.99 | PG | 24.49 |
| GE | 21.89 | SBC | 14.87 |
| GM | 9.86 | UTX | 14.87 |
| HD | 20.26 | WMT | 27.84 |
| HON | 23.36 | DIS | 37.10 |
| Mean = 23.25, variance = 90.13, standard deviation = 9.49 | | | |

*Source:* **www.stockselector.com**.

Table D-1 gives data on the *price to earnings ratio*—the famous P/E ratio—for 28 companies listed on the New York Stock Exchange (NYSE) for February 2, 2004 (at about 3 p.m.).[2] Assume that this is a random sample from the universe (population) of stocks listed on the NYSE, some 3000 or so. The P/E ratio of 27.96 for Alcoa (AA) listed in this table, for example, means that on that day the stock was selling at about 28 times its annual earnings. The P/E ratio is one of the key indicators for investors in the stock market.

Suppose our primary interest is not in any single P/E ratio, but in the *average* P/E ratio in the entire population of the NYSE listed stocks. Since we can obtain data on the P/E ratios of all the stocks listed on the NYSE, in principle, we can easily compute the average P/E ratio. In practice, that would be time-consuming and expensive. Could we use the data given in Table D-1 to compute the *average* P/E ratio of the 28 companies listed in this table and use this (sample) average as an *estimate* of the average P/E ratio in the entire population of the stocks listed on the NYSE? Specifically, if we let $X$ = P/E ratio of a stock and $\overline{X}$ = the average P/E ratio of the 28 stocks given in Table D-1, can we tell what the expected P/E ratio, $E(X)$, is in the NYSE population as a whole? *This process of generalizing from the sample value (e.g., $\overline{X}$) to the population value* (e.g., E[$X$]) *is the essence of statistical inference.* We will now discuss this topic in some detail.

[2]Since the price of the stock varies from day to day, the P/E ratio will vary from day to day, even though the earnings do not change. The stocks given in this table are members of the so-called Dow 30. In reality stock prices change very frequently when the stock market is open, but most newspapers quote the P/E ratios as of the end of the business day.

## D.2 ESTIMATION AND HYPOTHESIS TESTING: TWIN BRANCHES OF STATISTICAL INFERENCE

From the preceding discussion it can be seen that statistical inference proceeds along the following lines. There is some population of interest, say, the stocks listed on the NYSE, and we are interested in studying some aspect of this population, say, the P/E ratio. Of course, we may not want to study each and every P/E ratio, but only the average P/E ratio. Since collecting information on all the NYSE P/E ratios needed to compute the average P/E ratio is expensive and time-consuming, we may obtain a *random sample* of only a few stocks to get the P/E ratio of each of these sampled stocks and compute the sample average P/E ratio, say, $\overline{X}$. $\overline{X}$ is an **estimator,** also known as a (sample) *statistic*, of the population average P/E ratio, $E(X)$, which is called the (population) **parameter.** (Refer to the discussion in Appendix B). For example, the mean and variance are the parameters of the normal distribution. A particular numerical value of the estimator is called an *estimate* (e.g., an $\overline{X}$ value of 23). Thus, *estimation is the first step in statistical inference*. Having obtained an estimate of a parameter, we next need to find out how good that estimate is, for an estimate is not likely to equal the true parameter value. If we obtain two or more random samples of 28 stocks each and compute $\overline{X}$ for each of these samples, the two estimates will probably not be the same. This variation in estimates from sample to sample is known as *sampling variation* or *sampling error.*[3] Are there any criteria by which we can judge the "goodness" of an estimator? In Section D.4 we discuss some of the commonly used criteria to judge the goodness of an estimator.

Whereas estimation is one side of statistical inference, hypothesis testing is the other. In *hypothesis testing* we may have prior judgment or expectation about what value a particular parameter may assume. For example, prior knowledge or an expert opinion tells us that the true average P/E ratio in the population of NYSE stocks is, say, 20. Suppose a particular random sample of 28 stocks gives this estimate as 23. Is this value of 23 close to the *hypothesized value* of 20? Obviously, the number 23 is different from the number 20. But the important question here is this: Is 23 *statistically different* from 20? We know that because of sampling variation there is likely to be a difference between a (sample) estimate and its population value. It is possible that statistically the number 23 may not be very different from the number 20, in which case we may not reject the hypothesis that the true average P/E ratio is 20. But how do we decide that? This is the essence of the topic of *hypothesis testing*, which we will discuss in Section D.5.

With these preliminaries, let us examine the twin topics of estimation and hypothesis testing in some detail.

[3]Notice that this sampling error is not deliberate, but it occurs because we have a random sample and the elements included in the sample will vary from sample to sample. This is inevitable in any analysis based on a sample.

## D.3 ESTIMATION OF PARAMETERS

In Appendix C we considered several theoretical probability distributions. Often we know or are willing to assume that a random variable $X$ follows a particular distribution, but we do not know the value(s) of the parameter(s) of the distribution. For example, if $X$ follows the normal distribution, we may want to know the values of its two parameters, namely, the mean $E(X) = \mu_X$ and the variance $\sigma_X^2$. To estimate these unknowns, the usual procedure is to assume that we have a *random sample* of size $n$ from the known probability distribution and to use the sample to estimate the unknown parameters. Thus, we can use the sample mean as an estimate of the population mean (or expected value) and the sample variance as an estimate of the population variance. This procedure is known as the *problem of estimation*. The problem of estimation can be broken down into two categories: point estimation and interval estimation.

To fix the ideas, assume that the random variable (r.v.), $X$ (P/E ratio), is normally distributed with a certain mean and a certain variance, but for now we do not know the values of these parameters. Suppose, however, we have a random sample of 28 P/E ratios (28 $X$'s) from this normal population, as shown in Table D-1.

How can we use these sample data to compute the population mean value $\mu_X = E(X)$ and the population variance $\sigma_X^2$? More specifically, suppose our immediate interest is in finding out $\mu_X$.[4] How do we go about it? An obvious choice is the sample mean $\overline{X}$ of the 28 P/E ratios shown in Table D-1, which is 23.25. We call this *single numerical value* the **point estimate** of $\mu_X$, and the formula $\overline{X} = \sum_1^{28} X_i/n$ that we used to compute this point estimate is called the *point estimator*, or *statistic*. Notice that a *point estimator, or a statistic, is an r.v., as its value will vary from sample to sample.* (Recall our sampling experiment in Example C-6.) Therefore, how reliable is a specific estimate such as 23.25 of the true $\mu_X$? In other words, how can we rely on just one estimate of the true population mean? Would it not be better to state that although $\overline{X}$ is the single best guess of the true population mean, the interval, say, from 19 to 24, most likely includes the true $\mu_X$? This is essentially the idea behind **interval estimation.** We will now consider the actual mechanics of obtaining interval estimates.

The key idea underlying interval estimation is the notion of **sampling,** or **probability, distribution** of an estimator such as the sample mean $\overline{X}$, which we have already discussed in Appendix C. In Appendix C we saw that if an r.v. $X \sim N(\mu_X, \sigma_X^2)$, then

$$\overline{X} \sim \left(\mu_X, \frac{\sigma_X^2}{n}\right) \tag{D.1}$$

or

$$Z = \frac{(\overline{X} - \mu_X)}{\sigma_X/\sqrt{n}} \sim N(0, 1) \tag{D.2}$$

[4]This discussion can be easily extended to estimate $\sigma_X^2$.

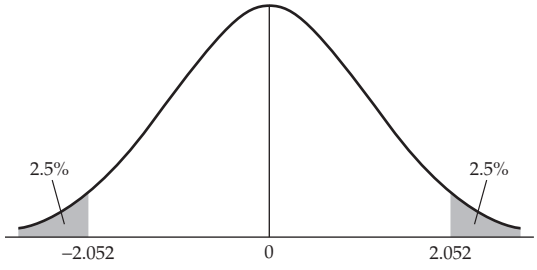


**FIGURE D-1** The *t* distribution for 27 d.f.

That is, the sampling distribution of the sample mean $\overline{X}$ also follows the normal distribution with the stated parameters.[5]

As pointed out in Appendix C, $\sigma_X^2$ is not generally known, but if we use its estimator $S_x^2 = \sum(X_i - \overline{X})^2/(n-1)$, then we know that

$$t = \frac{(\overline{X} - \mu_X)}{S_x/\sqrt{n}} \tag{D.3}$$

follows the $t$ distribution with $(n-1)$ degrees of freedom (d.f.).

To see how Equation (D.3) helps us to obtain an interval estimation of the $\mu_X$ of our P/E example, note that we have a total of 28 observations and, therefore, 27 d.f. Now if we consult the $t$ table (Table E-2) given in Appendix E, we notice that for 27 d.f.,

$$P(-2.052 \leq t \leq 2.052) = 0.95 \tag{D.4}$$

as shown in Figure D-1. That is, for 27 d.f., the probability is 0.95 (or 95 percent) that the interval (−2.052, 2.052) will include the $t$ value computed from Eq. (D.3).[6] These $t$ values, as we will see shortly, are known as **critical *t* values;** they show what percentage of the area under the $t$ distribution curve (see Figure D-1) lies between those values (note that the total area under the curve is 1); $t = -2.052$ is called the *lower critical t value* and $t = 2.052$ is called the *upper critical t value*.

Now substituting the $t$ value from Eq. (D.3) into Eq. (D.4), we obtain

$$P\left(-2.052 \leq \frac{(\overline{X} - \mu_X)}{S_x/\sqrt{n}} \leq 2.052\right) \tag{D.5}$$

Simple algebraic manipulation will show that Equation (D.5) can be expressed *equivalently* as

$$P\left(\overline{X} - 2.052\frac{S_x}{\sqrt{n}} \leq \mu_X \leq \overline{X} + 2.052\frac{S_x}{\sqrt{n}}\right) = 0.95 \tag{D.6}$$

[5]Note that if $X$ does not follow the normal distribution, $\overline{X}$ will follow the normal distribution à la the central limit theorem if $n$, the sample size, is sufficiently large.

[6]Needless to say, these values will depend on the d.f. as well as on the level of probability used. For example, for the same d.f. $P(-2.771 \leq t \leq 2.771) = 0.99$.

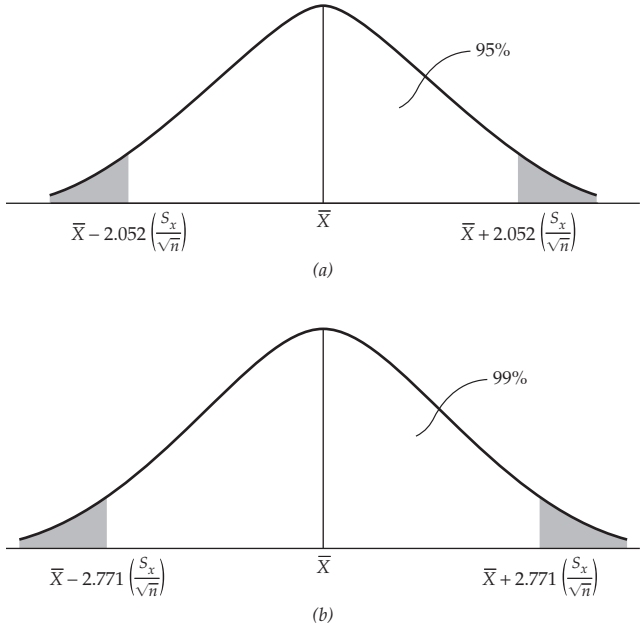


**FIGURE D-2** (*a*) 95% and (*b*) 99% confidence intervals for $\mu_X$ for 27 d.f.

Equation (D.6) provides an *interval estimator* of the true $\mu_X$.

In statistics we call Eq. (D.6) a 95% **confidence interval (CI)** for the true but unknown population mean $\mu_X$ and 0.95 is called the **confidence coefficient**. In words, Eq. (D.6) says that the *probability is 0.95 that the random interval* $(\bar{X} \pm 2.052S_x/\sqrt{n})$ *contains the true* $\mu_X$. $(\bar{X} - 2.052S_x/\sqrt{n})$ is called the **lower limit** of the interval and $(\bar{X} + 2.0096S_x/\sqrt{n})$ is the **upper limit** of the interval. See Figure D-2.

Before proceeding further, note this important point: The interval given in Eq. (D.6) is a **random interval** because it is based on $\bar{X}$ and $S_x/\sqrt{n}$, which will vary from sample to sample. The true or population mean $\mu_X$, although unknown, is some fixed number and therefore is not random. Thus, *one should not say that the probability is 0.95 that* $\mu_X$ *lies in this interval. The correct statement, as noted earlier, is that the probability is 0.95 that the random interval, Eq. (D.6), contains the true* $\mu_X$. *In short, the interval is random and not the parameter* $\mu_X$.

Returning to our P/E example of Table D-1, we have $n = 28$, $\bar{X} = 23.25$, and $S_x = 9.49$. Plugging these values into Eq. (D.6), we obtain

$$23.25 - \frac{(2.052)(9.49)}{\sqrt{28}} \leq \mu_X \leq 23.25 + \frac{(2.052)(9.49)}{\sqrt{28}}$$

which yields

$$19.57 \leq \mu_X \leq 26.93 \ (approx) \qquad \textbf{(D.7)}$$

as the 95% confidence interval for $\mu_X$.

Equation (D.7) says, in effect, that if we construct intervals like Eq. (D.7), say, 100 times, then 95 out of 100 such intervals will include the true $\mu_X$.[7] Incidentally, note that for our P/E example the lower limit of the interval is 19.57 and the upper limit is 26.93.

Thus, *interval estimation, in contrast to point estimation (such as 23.25), provides a range of values that will include the true value with a certain degree of confidence or probability (such as 0.95).* If we have to give one best estimate of the true mean, it is the point estimate 23.25, but if we want to be less precise we can give the interval (19.57 to 26.93) as the range that most probably includes the true mean value with a certain degree of confidence (95 percent in the present instance).

More generally, suppose $X$ is an r.v. with some probability distribution function (PDF). Suppose further that we want to estimate a parameter of this distribution, say, its mean value $\mu_X$. Toward that end, we obtain a random sample of $n$ values, $X_1, X_2, \cdots, X_n$, and compute two statistics (or estimators) $L$ and $U$ from this sample such that

$$P(L \leq \mu_X \leq U) = 1 - \alpha \qquad 0 < \alpha < 1 \tag{D.8}$$

That is, the probability is $(1 - \alpha)$ that the *random interval* from $L$ to $U$ contains the true $\mu_X$. $L$ is called the lower limit of the interval and $U$ is called the upper limit. This interval is known as a *confidence interval* of size $(1 - \alpha)$ for $\mu_X$ (or any parameter for that matter), and $(1 - \alpha)$ is known as the *confidence coefficient.* If $\alpha = 0.05$, $(1 - \alpha) = 0.95$, meaning that if we construct a confidence interval with a confidence coefficient of 0.95, then in repeated such constructions, 95 out of 100 intervals can be expected to include the true $\mu_X$. In practice, $(1 - \alpha)$ is often multiplied by 100 to express it in percent form (e.g., 95 percent). In statistics alpha ($\alpha$) is known as the **level of significance,** or, alternatively, **the probability of committing a type I error**, which is defined and discussed in Section D.5.

Now that we have seen how to establish confidence intervals, what do we do with them? As we will see in Section D.5, confidence intervals make our task of testing hypotheses—the twin of statistical inference—much easier.

## D.4 PROPERTIES OF POINT ESTIMATORS

In the P/E example we used the sample mean $\overline{X}$ as a point estimator of $\mu_X$, as well as to obtain an interval estimator of $\mu_X$. But why did we use $\overline{X}$? It is well

[7] Be careful again. We cannot say that the probability is 0.95 that the particular interval in Eq. (D.7) includes the true $\mu_X$; it may or may not. Therefore, statements like $P(19.5 \leq \mu_X \leq 26.93) = 0.95$ *are not permissible under the classical approach to hypothesis testing.* Intervals like those in Eq. (D.7) are to be interpreted in the repeated sampling sense that if we construct such intervals a large number of times, then 95 percent of such intervals will include the true mean value; the particular interval in Eq. (D.7) is just one realization of the interval estimator in Eq. (D.6).

known that besides the sample mean, the (sample) median or the (sample) mode also can be used as point estimators of $\mu_X$.[8]

In practice, the sample mean is the most frequently used measure of the population mean because it satisfies several properties that statisticians deem desirable. Some of these properties are:

1. Linearity
2. Unbiasedness
3. Minimum variance
4. Efficiency
5. Best linear unbiased estimator (BLUE)
6. Consistency

We will now discuss these properties somewhat heuristically.

### Linearity

*An estimator is said to be a* **linear estimator** *if it is a linear function of the sample observations.* The sample mean is obviously a linear estimator because

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

is a linear function of the observations, the $X$'s. (*Note:* The $X$'s appear with an index or power of 1 only.)

In statistics a linear estimator is generally much easier to deal with than a nonlinear estimator.

### Unbiasedness

If there are several estimators of a population parameter (i.e., several methods of estimating that parameter), and if one or more of these estimators *on the average* coincide with the true value of the parameter, we say that such estimators are **unbiased estimators** of that parameter. Put differently, if in repeated applications of a method the *mean* value of the estimators coincides with the true parameter value, that estimator is called an unbiased estimator. More formally, an estimator, say, $\overline{X}$, is an unbiased estimator of $\mu_X$ if

$$E(\overline{X}) = \mu_X \tag{D.9}$$

[8]The *median* is that value of a random variable that divides the total PDF into two halves such that half the values in the population exceed it and half are below it. To compute the median from a sample, arrange the observations in increasing order; the median is the middle value in this order. For example, if we have observations 7, 3, 6, 11, 5 and rearrange them in increasing order, we obtain 3, 5, 6, 7, 11. The median, or the middlemost value, here is 6. The *mode* is the most popular or frequent value of the random variable. For example, if we have observations 3, 5, 7, 5, 8, 5, 9, the *modal value* is 5 since it occurs most frequently.

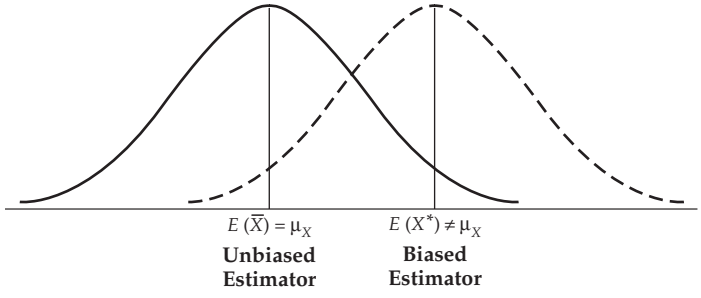


**FIGURE D-3** Biased ($X^*$) and unbiased ($\overline{X}$) estimators of population mean value, $\mu_X$

as shown in Figure D-3. If this is not the case, however, then we call that estimator a *biased estimator,* such as the estimator $X^*$ shown in Figure D-3.

**Example D.1.**

Let $X_i \sim N(\mu_X, \sigma_X^2)$, then, as we saw in Appendix C. $\overline{X}$, based on a random sample of size $n$ from this population, is distributed with mean $E(\overline{X}) = \mu_X$ and var $(\overline{X}) = \sigma_X^2/n$. Thus, the sample mean $\overline{X}$ is an unbiased estimator of true $\mu_X$. If we draw repeated samples of size $n$ from this normal population and compute $\overline{X}$ for each sample, then on the average, $\overline{X}$ will coincide with $\mu_X$. But notice carefully that we cannot say that in a single sample, such as the one in Table D-1, the computed mean of 23.25 will necessarily coincide with the true mean value.

**Example D.2.**

Again, let $X_i \sim N(\mu_X, \sigma_X^2)$, and suppose we draw a random sample of size $n$ from this population. Let $X_{\text{med}}$ represent the median value of this sample. It can be shown that $E(X_{\text{med}}) = \mu_X$. In words, the median from this population is also an unbiased estimator of the true mean. Notice also that unbiasedness is a *repeated sampling property;* that is, if we draw several samples of size $n$ from this population and compute the median value for each sample, then the average of the median values obtained will tend to approach $\mu_X$.

### Minimum Variance

Figure D-4 shows the sampling distributions of three estimators of $\mu_X$, obtained from three different estimators, $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\mu}_3$.

Now an estimator of, say, $\mu_X$, is said to be a **minimum-variance** estimator if its variance is smaller than any other estimator of $\mu_X$. As you can see from Fig. D-4, the variance of $\hat{\mu}_3$ is the smallest of the three estimators shown there. Hence, it is a minimum-variance estimator. But note that $\hat{\mu}_3$ is a biased estimator. (Why?)

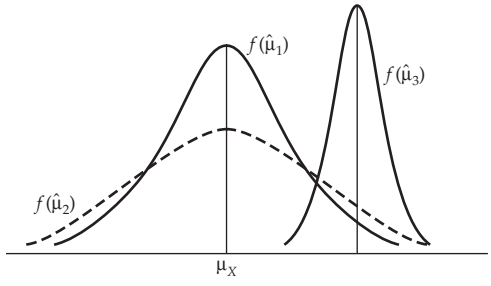


**FIGURE D-4** Distribution of three estimators of $\mu_X$

## Efficiency

The property of unbiasedness, although desirable, is not adequate by itself. What happens if we have two or more estimators of a parameter and they are all unbiased? How do we choose among them?

Suppose we have a random sample of $n$ values of an r.v. $X$ such that each $X \sim N(\mu_X, \sigma_X^2)$. Let $\overline{X}$ and $X_{\text{med}}$ be the mean and median values obtained from this sample. We already know that

$$\overline{X} \sim N(\mu_X, \sigma^2/n) \tag{D.10}$$

It can also be shown that *if the sample size is large,*

$$X_{\text{med}} \sim N(\mu_X, (\pi/2)(\sigma^2/n)) \tag{D.11}$$

where $\pi = 3.142$ (approx.). That is, in large samples, the median computed from a random sample of a normal population also follows a normal distribution with the same mean $\mu_X$ but with a variance that is larger than the variance of $\overline{X}$ by the factor $\pi/2$, which can be visualized from Figure D-5. As a matter of fact, by forming the ratio

$$\frac{\text{var}\,(X_{\text{med}})}{\text{var}\,(\overline{X})} = \frac{\pi}{2}\frac{\sigma^2/n}{\sigma^2/n} = \frac{\pi}{2} = 1.571 \quad \text{(approx)} \tag{D.12}$$

we show that the variance of the sample median is ≈57 percent larger than the variance of the sample mean.

Now given Figure D-5 and the preceding discussion, which estimator would you choose? Common sense suggests that we choose $\overline{X}$ over $X_{\text{med}}$, for although both estimators are unbiased, $\overline{X}$ has a smaller variance than $X_{\text{med}}$. Therefore if we use $\overline{X}$ in repeated sampling, we will estimate $\mu_X$ more accurately than if we were to use the sample median. In short, $\overline{X}$ provides a more *precise* estimate of the population mean than the median $X_{\text{med}}$. In statistical language we say that $\overline{X}$ is an **efficient estimator.** Stated more formally, *if we consider only unbiased estimators of a parameter, the one with the smallest variance is called the best, or efficient, estimator.*

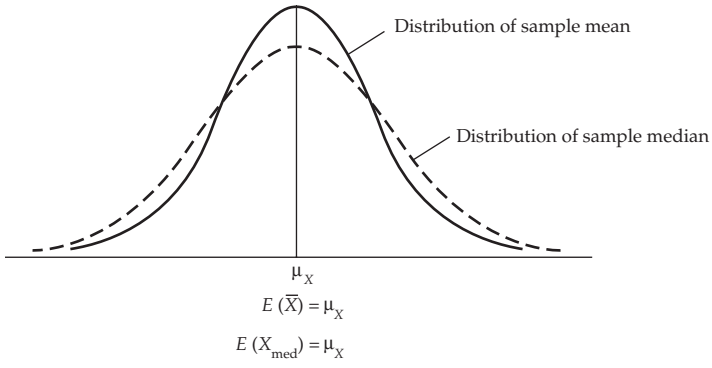


**FIGURE D-5** An example of an efficient estimator (sample mean)

### Best Linear Unbiased Estimator (BLUE)

In econometrics the property that is frequently encountered is the property **best linear unbiased estimator,** or **BLUE** for short. *If an estimator is linear, is unbiased, and has minimum variance in the class of all linear unbiased estimators of a parameter, it is called a best linear unbiased estimator.* Obviously, this property combines the properties of linearity, unbiasedness, and minimum variance. In Chapters 3 and 4 we will see the importance of this property.

### Consistency

To explain the property of consistency, suppose $X \sim N(\mu_X, \sigma_X^2)$ and we draw a random sample of size $n$ from this population. Now consider two estimators of $\mu_X$.

$$\overline{X} = \sum \frac{X_i}{n} \tag{D.13}$$

$$X^* = \sum \frac{X_i}{n+1} \tag{D.14}$$

The first estimator is the usual sample mean. Now, as we already know

$$E(\overline{X}) = \mu_X$$

and it can be shown that

$$E(X^*) = \left(\frac{n}{n+1}\right)\mu_X \tag{D.15}$$

Since $E(X^*)$ is not equal to $\mu_X$, $X^*$ is obviously a *biased estimator.* (For proof, see Problem D. 21.)

But suppose we increase the sample size. What would you expect? The estimators $\overline{X}$ and $X^*$ differ only in that the former has $n$ in the denominator whereas the latter has $(n + 1)$. But as the sample increases, we should not find

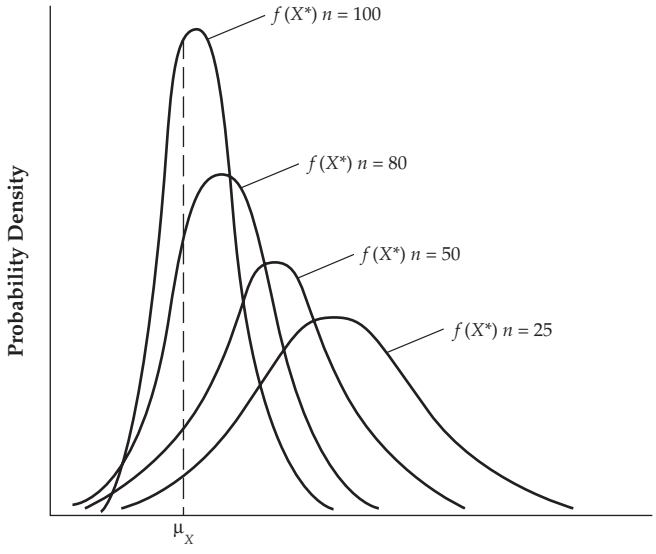


**FIGURE D-6** The property of consistency. The behavior of the estimator $X^*$ of population mean $\mu_X$ as the sample size increases

much difference between the two estimators. That is, as the sample size increases, $X^*$ also will approach the true $\mu_X$. In statistics such an estimator is known as a **consistent estimator.** Stated more formally, *an estimator (e.g., $X^*$) is said to be a consistent estimator if it approaches the true value of the parameter as the sample size gets larger and larger.* As we will see in the main chapters of the text, sometimes we may not be able to obtain an unbiased estimator, but we can obtain a consistent estimator.[9] The property of consistency is depicted in Figure D-6.

## D.5 STATISTICAL INFERENCE: HYPOTHESIS TESTING

Having studied in some detail the estimation branch of statistical inference, we will now consider its twin, hypothesis testing. Although the general nature of hypothesis testing was discussed earlier, we study it here in some detail.

Let us return to the P/E example given in Table D-1. In Section D.3, based on a random sample of 28 P/E ratios, we established a 95% confidence interval for $\mu_X$, the true but unknown average P/E ratio in the population of the stocks listed on the NYSE. Now let us reverse our strategy. Instead of establishing a

[9]Note the critical difference between an unbiased and a consistent estimator. If we fix the sample size and draw several random samples of an r.v. from some probability distribution to estimate a parameter of this distribution, then unbiasedness requires that *on the average* we should be able to obtain the true parameter value. In establishing consistency, on the other hand, we see the behavior of an estimator as the sample size increases. If a sample size is reasonably large and the estimator based on that sample size approaches the true parameter value, then that estimator is a consistent estimator.

confidence interval, suppose we *hypothesize* that the true $\mu_X$ takes a particular numerical value (e.g., $\mu_X = 18.5$). Our task now is to test this hypothesis.[10] How do we test this hypothesis—that is, support or refute it?

In the language of hypothesis testing a hypothesis such as $\mu_X = 18.5$ is called a **null hypothesis** and is generally denoted by the symbol $H_0$. Thus, $H_0$: $\mu_X = 18.5$. The null hypothesis is usually tested against an **alternative hypothesis**, denoted by the symbol $H_1$. The alternative hypothesis can take one of these forms:

$H_1$:$\mu_X > 18.5$, which is called a **one-sided** or **one-tailed** alternative hypothesis, or

$H_1$:$\mu_X < 18.5$, also a **one-sided** or **one-tailed** alternative hypothesis, or

$H_1$:$\mu_X \neq 18.5$, which is called a **composite, two-sided,** or **two-tailed** alternative hypothesis. That is, the true mean value is either greater than or less than 18.5.[11]

To test the null hypothesis (against the alternative hypothesis), we use the sample data (e.g., the sample average P/E ratio of 23.25 obtained from the sample in Table D-1) and statistical theory to develop decision rules that will tell us whether the sample evidence supports the null hypothesis. If the sample evidence supports the null hypothesis, we do not reject $H_0$, but if it does not, we reject $H_0$. In the latter case we may accept the alternative hypothesis, $H_1$.

How do we develop these decision rules? There are two complementary approaches: (1) confidence interval and (2) test of significance. We illustrate each with the aid of our P/E example. Assume that

$$H_0: \mu_X = 18.5$$

$$H_1: \mu_X \neq 18.5 \quad (\textit{a two-sided hypothesis})$$

### The Confidence Interval Approach to Hypothesis Testing

To test the null hypothesis, suppose we have the sample data given in Table D-1. From these data we computed the sample mean of 23.25. We know from our discussion in Section D.3 that the sample mean is distributed normally with mean $\mu_X$ and variance $\sigma_X^2/n$. But since the true variance is unknown, we replace it with the sample variance, in which case we know that the sample mean follows the $t$ distribution, as shown in Eq. (D.3). Based on the $t$ distribution, we obtain the following 95% confidence interval for:

$$19.57 \leq \mu_X \leq 26.93 \qquad \textbf{(D.16) = (D.7)}$$

We know that confidence intervals provide a range of values that may include the true $\mu_X$ with a certain degree of confidence, such as 95 percent. Therefore, if

[10] A *hypothesis* is "something considered to be true for the purpose of investigation or argument" (*Webster's*), or a "supposition made as a basis for reasoning, or as a starting point for further investigation from known facts" (*Oxford English Dictionary*).

[11] There are various ways of stating the null and alternative hypotheses. For example, we could have $H_0$:$\mu_X \geq 13$ and $H_1$: $\mu_X < 13$.

this interval does not include a particular null hypothesized value such as $\mu_X = 18.5$, could we not reject this null hypothesis? Yes, we can, with 95% confidence.

From the preceding discussion it should be clear that the topics of confidence interval and hypothesis testing are intimately related. In the language of hypothesis testing, the 95% confidence interval shown in inequality (D.7) (see Fig. D-2) is called the **acceptance region** and the area outside the acceptance region is called the **critical region,** or **the region of rejection,** of the null hypothesis. The lower and upper limits of the acceptance region are called **critical values.** In this language, if the acceptance region includes the value of the parameter under $H_0$, we do not reject the null hypothesis. But if it falls outside the acceptance region (i.e., it lies within the rejection region), we reject the null hypothesis. In our example we reject the null hypothesis that $\mu_X = 18.5$ since the acceptance region given in Eq. (D.7) does not include the null-hypothesized value. It should be clear now why the boundaries of the acceptance region are called critical values, for they are the dividing line between accepting and rejecting a null hypothesis.

### Type I and Type II Errors: A Digression

In our P/E example we rejected $H_0$:$\mu_X = 18.5$ because our sample evidence of $\overline{X} = 23.25$ does not seem to be compatible with this hypothesis. Does this mean that the sample shown in Table D-1 did not come from a normal population whose mean value was 18.5? We cannot be absolutely sure, for the confidence interval given in inequality (D.7) is 95 and not 100 percent. If that is the case, we would be making an error in rejecting $H_0$:$\mu_X = 18.5$. In this case we are said to commit a **type I error,** that is, *the error of rejecting a hypothesis when it is true.* By the same token, suppose $H_0$:$\mu_X = 21$, in which case, as inequality (D.7) shows, we would not reject this null hypothesis. But quite possibly the sample in Table D-1 did not come from a normal distribution with a mean value of 21. Thus, we are said to commit a **type II error,** that is, *the error of accepting a false hypothesis.* Schematically,

| | Reject $H_0$ | Do not reject $H_0$ |
|---|---|---|
| $H_0$ is true | Type I error | Correct decision |
| $H_0$ is false | Correct decision | Type II error |

Ideally, we would like to minimize both these errors. But, unfortunately, for *any given sample size*,[12] it is not possible to minimize both errors simultaneously. The classical approach to this problem, embodied in the work of statisticians Neyman and Pearson, is to assume that a type I error is likely to be more serious in practice than a type II error. Therefore, we should try to keep the probability

[12]The only way to decrease a type II error without increasing a type I error is to increase the sample size, which may not always be easy.

of committing a type I error at a fairly low level, such as 0.01 or 0.05, and then try to minimize a type II error as much as possible.[13]

In the literature the probability of committing a type I error is designated as $\alpha$ and is called the **level of significance,**[14] and the probability of committing a type II error is designated as $\beta$. Symbolically,

$$\text{Type I error} = \alpha = \text{prob. } (\textit{rejecting } H_0 \mid H_0 \textit{ is true})$$

$$\text{Type II error} = \beta = \text{prob. } (\textit{accepting } H_0 \mid H_0 \textit{ is false})$$

The probability of *not* committing a type II error, that is, rejecting $H_0$ when it is false, is $(1 - \beta)$, which is called the **power of the test**.

The standard, or classical, approach to hypothesis testing is to fix $\alpha$ at levels such as 0.01 or 0.05 and then try to maximize the power of the test; that is, to minimize $\beta$. How this is actually accomplished is involved, and so we leave the subject for the references.[15] Suffice it to note that, in practice, the classical approach simply specifies the value of $\alpha$ without worrying too much about $\beta$. But keep in mind that, in practice, in making a decision there is a trade-off between the significance level and the power of the test. That is, for a *given sample size*, if we try to reduce the probability of a type I error, we ipso facto increase the probability of a type II error and therefore reduce the power of the test. Thus, instead of using $\alpha = 5$ percent, if we were to use $\alpha = 1$ percent, we may be very confident when we reject $H_0$, but we may not be so confident when we do not reject it.

Since the precedent point is important, let us illustrate. For our P/E ratio example, in Eq. (D.7) we established a 95% confidence. Let us still assume that $H_0$:$\mu_X = 18.5$ but now fix $\alpha = 1$ percent and obtain the 99% confidence interval, which is (noting that for 99% CI, the critical $t$ values are $(-2.771, 2.771)$ for 27 d.f.):

$$18.28 \leq \mu_X \leq 28.22 \qquad \textbf{(D.17)}$$

This 99% confidence interval is also shown in Fig. D-2. Obviously, this interval is wider than the 95% confidence interval. Since this interval includes the hypothesized value of 18.5, we do not reject the null hypothesis, whereas in Eq. (D.7) we rejected the null hypothesis on the basis of a 95% confidence interval. What now? By reducing a type I error from 5 percent to 1 percent, we have increased the probability of a type II error. That is, in not rejecting the null hypothesis on the basis of Eq. (D.17), we may be falsely accepting the hypothesis

[13] To Bayesian statisticians this procedure sounds rather arbitrary because it does not consider carefully the relative seriousness of the two types of errors. For further discussion of this and related points, see Robert L. Winkler, *Introduction to Bayesian Inference and Decision,* Holt, Rinehart and Winston, New York, 1972, Chap. 7.

[14] $\alpha$ is also known as the *size of the (statistical) test.*

[15] For a somewhat intuitive discussion of this topic, see Gujarati and Porter, *Basic Econometrics,* 5th ed., McGraw-Hill, New York, 2009, pp. 833–835. Statistical packages, such as MINITAB, can calculate the power of a test of size $\alpha$.

that the true $\mu_X$ is 18.5. So, always keep in mind the trade-off involved between type I and type II errors.

You will recognize that the **confidence coefficient** $(1 - \alpha)$ discussed earlier is simply 1 minus the probability of committing a type I error. Thus, a 95% confidence coefficient means that we are prepared to accept at most a 5 percent probability of committing a type I error—we do not want to reject the true hypothesis by more than 5 out of 100 times. *In short, a 5% level of significance or a 95% level or degree of confidence means the same thing.*

Let us consider another example to illustrate further the confidence interval approach to hypothesis testing.

**Example D.3.**

The number of peanuts contained in a jar follows the normal distribution, but we do not know its mean and standard deviation, both of which are measured in ounces. Twenty jars were selected randomly and it was found that the sample mean was 6.5 ounces and the sample standard deviation was 2 ounces. Test the hypothesis that the true mean value is 7.5 ounces against the hypothesis that it is different from 7.5. Use $\alpha = 1\%$.

***Answer:*** Letting $X$ denote the number of peanuts in a jar, we are given that $X \sim N(\mu_X, \sigma_X^2)$, both parameters being unknown. Since the true variance is unknown, if we use its estimator $S_x^2$, it follows that

$$t = \frac{\overline{X} - \mu_X}{S_x/\sqrt{n}} \sim t_{19}$$

That is, the $t$ distribution with 19 d.f.

From the $t$ distribution table given in Table E-2 in Appendix E, we observe that for 19 d.f.,

$$P(-2.861 \leq t \leq 2.861) = 0.99$$

Then from expression (D.6) we obtain

$$P\left(\overline{X} - 2.861\frac{S_x}{\sqrt{20}} \leq \mu_X \leq \overline{X} + 2.861\frac{S_x}{\sqrt{20}}\right) = 0.99$$

Substituting $\overline{X} = 6.5$, $S_x = 2$, and $n = 20$ into this inequality, we obtain

$$5.22 \leq \mu_X \leq 7.78 \quad \text{(approx.)} \tag{D.18}$$

as the 99% confidence interval for $\mu_X$. Since this interval includes the hypothesized value of 7.5, we do not reject the null hypothesis that the true $\mu_X = 7.5$.

The null hypothesis in our P/E example was $\mu_X = 18.5$ and the alternative hypothesis was that $\mu_X \neq 18.5$, which is a two-sided, or composite, hypothesis.

How do we handle one-sided alternative hypotheses such as $\mu_X < 18.5$ *or* $\mu_X > 18.5$? Although the confidence interval approach can be easily adapted to construct one-sided confidence intervals, in practice it is much easier to use the test of significance approach to hypothesis testing, which we will now discuss.

### The Test of Significance Approach to Hypothesis Testing

The test of significance is an alternative, but complementary and perhaps shorter, approach to hypothesis testing. To see the essential points involved, return to the P/E example and Eq. (D.3). We know that

$$t = \frac{\overline{X} - \mu_X}{S_x/\sqrt{n}} \qquad \textbf{(D.19) = (D.3)}$$

follows the $t$ distribution with $(n - 1)$ d.f. In any concrete application we will know the values of $\overline{X}$, $S_x$, and $n$. The only unknown value is $\mu_X$. But if we specify a value for $\mu_X$, as we do under $H_0$, then the right-hand side of Eq. (D.3) is known, and therefore we will have a unique $t$ value. And since we know that the $t$ of Eq. (D.3) follows the $t$ distribution with $(n - 1)$, we simply look up the $t$ table to find out the probability of obtaining such a $t$ value.

Observe that if the difference between $\overline{X}$ and $\mu_X$ is small (in absolute terms), then, as Eq. (D.3) shows, the $|t|$ value will also be small, where $|t|$ means the absolute $t$ value. In the event that $\overline{X} = \mu_X$, $t$ will be zero, in which case we do not reject the null hypothesis. *Therefore, as the $|t|$ value increasingly deviates from zero, we will tend to reject the null hypothesis.* As the $t$ table shows, for any given d.f., the probability of obtaining an increasingly higher $|t|$ value becomes progressively smaller. *Thus, as $|t|$ gets larger, we will be more and more inclined to reject the null hypothesis.* But how large must $|t|$ be before we can reject the null hypothesis? The answer, as you would suspect, depends on $\alpha$, the probability of committing a type I error, as well as on the d.f., as we will demonstrate shortly.

This is the general idea behind the test of significance approach to hypothesis testing. The key idea here is the **test statistic**—the ***t* statistic**—and its probability distribution under the hypothesized value of $\mu_X$. Appropriately, in the present instance the test is known as the ***t* test** since we use the $t$ distribution. (For details of the $t$ distribution, see Section C.2).

In our P/E example $\overline{X} = 23.25$, $S_x = 9.49$ and $n = 28$. Let $H_0$:$\mu_X = 18.5$ and $H_1$:$\mu_X \neq 18.5$, as before. Therefore,

$$t = \frac{23.25 - 18.5}{9.49/\sqrt{28}} = 2.6486 \qquad \textbf{(D.20)}$$

Is the computed $t$ value such that we can reject the null hypothesis? We cannot answer this question without first specifying what chance we are willing to take if we reject the null hypothesis when it is true. In other words, to answer this question, we must specify $\alpha$, the probability of committing a type I error. Suppose we fix $\alpha$ at 5 percent. Since the alternative hypothesis is two-sided, we want to divide the risk of a type I error equally between the two tails of the

$t$ distribution—the two critical regions—so that if the computed $t$ value lies in either of the rejection regions, we can reject the null hypothesis.

Now for 27 d.f., as we saw earlier, the 5% *critical t values* are −2.052 and +2.052, as shown in Fig. D-1. The probability of obtaining a $t$ value equal to or smaller than −2.0096 is 2.5 percent and that of obtaining a $t$ value equal to or greater than +2.0096 is also 2.5 percent, giving the total probability of committing a type I error of 5 percent.

As Fig. D-1 also shows, the computed $t$ value for our example is about 2.6, which obviously lies in the right tail critical region of the $t$ distribution. We therefore reject the null hypothesis that the true average P/E ratio is 18.5. If that hypothesis were true, we would not have obtained a $t$ value as large as 2.6 (in absolute terms); the probability of our obtaining such a $t$ value is much smaller than 5 percent—our prechosen probability of committing a type I error. Actually, the probability is much smaller than 2.5 percent. (Why?)

In the language of the test of significance we frequently come across the following two terms:

1. A test (statistic) is *statistically significant.*
2. A test (statistic) is *statistically insignificant*.

When we say that a test is statistically significant, we generally mean that we can reject the null hypothesis. That is, the probability that the observed difference between the sample value and the hypothesized value is due to mere chance is small, less than $\alpha$ (the probability of a type I error). By the same token, when we say that a test is statistically insignificant, we do not reject the null hypothesis. In this case, the observed difference between the sample value and the hypothesized value could very well be due to sampling variation or due to mere chance (i.e., the probability of the difference is much greater than $\alpha$).

When we reject the null hypothesis, we say that our finding is *statistically significant.* On the other hand, when we do not reject the null hypothesis, we say that our finding is *not statistically significant.*

**One or Two-Tailed Test?** In all the examples considered so far the alternative hypothesis was two-sided, or two-tailed. Thus, if the average P/E ratio were equal to 18.5 under $H_0$, it was either greater than or less than 18.5 under $H_1$. In this case if the test statistic fell in either tail of the distribution (i.e., the rejection region), we rejected the null hypothesis, as is clear from Figure D-7(*a*).

However, there are occasions when the null and alternative hypotheses are one-sided, or one-tailed. For example, if for the P/E example we had $H_0$:$\mu_X \leq 18.5$ and $H_1$:$\mu_X > 18.5$, the alternative hypothesis is one-sided. How do we test this hypothesis?

The testing procedure is exactly the same as that used in previous cases except instead of finding out two critical values, we determine only a single critical value of the test statistic, as shown in Fig. D-7. As this figure illustrates, the probability of committing a type I error is now concentrated only in one tail of the probability distribution, $t$ in the present case. For 27 d.f. and $\alpha = 5$ percent, the $t$ table will

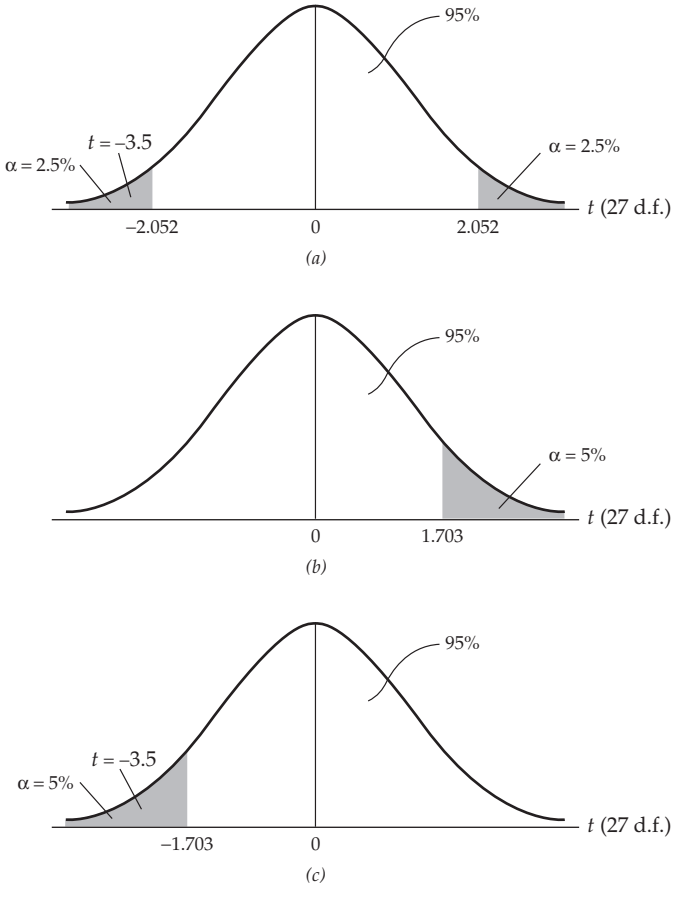


**FIGURE D-7** The *t* test of significance: (*a*) Two-tailed; (*b*) right-tailed; (*c*) left-tailed

show that the one-tailed critical *t* value is 1.703 (right tail) or −1.703 (left tail), as shown in Fig. D-7. For our P/E example, as noted before, the computed *t* value is about 2.43. Since the *t* value lies in the critical region of Fig. D-7(*b*), this *t* value is statistically significant. That is, we reject the null hypothesis that the true average P/E ratio is equal to (or less than) 18.5; the chances of that happening are much smaller than our prechosen probability of committing a type I error of 5 percent.

Table D-2 summarizes the *t* test of significance approach to testing the two-tailed and one-tailed null hypothesis.

In practice, whether we use the confidence interval approach or the test of significance approach to hypothesis testing is a matter of personal choice and convenience.

In the confidence interval approach we specify a plausible range of values (i.e., confidence interval) for the true parameter and find out if the confidence interval includes the hypothesized value of that parameter. If it does, we do not

**TABLE D-2** A SUMMARY OF THE *t* TEST

| Null hypothesis $H_0$ | Alternative hypothesis $H_1$ | Critical region Reject $H_0$ if |
|---|---|---|
| $\mu_X = \mu_0$ | $\mu_X > \mu_0$ | $t = \dfrac{\overline{X} - \mu_0}{S_x/\sqrt{n}} > t_{\alpha,d.f.}$ |
| $\mu_X = \mu_0$ | $\mu_X < \mu_0$ | $t = \dfrac{\overline{X} - \mu_0}{S_x/\sqrt{n}} < -t_{\alpha,d.f.}$ |
| $\mu_X = \mu_0$ | $\mu_X \neq \mu_0$ | $\lvert t\rvert = \dfrac{\overline{X} - \mu_0}{S_x/\sqrt{n}} > t_{\alpha/2,d.f.}$ |

*Note:* $\mu_0$ denotes the particular value of $\mu_X$ assumed under the null hypothesis. The first subscript on the *t* statistic shown in the last column is the level of significance, and the second subscript is the d.f. These are the critical *t* values.

reject that null hypothesis, but if it lies outside the confidence interval, we can reject the hypothesis.

In the test of significance approach, instead of specifying a range of plausible values for the unknown parameter, we pick a specific value of the parameter suggested by the null hypothesis; compute a test statistic, such as the *t* statistic; and find its sampling distribution and the probability of obtaining a specific value of such a test statistic. If this probability is very low, say, less than $\alpha = 5$ or 1 percent, we reject the particular null hypothesis. If this probability is greater than the preselected $\alpha$, we do not reject the null hypothesis.

**A Word about Accepting or Rejecting a Null Hypothesis** In this book we have used the terminology "reject" or "do not reject" a null hypothesis rather than "reject" or "accept" a hypothesis. This is in the same spirit as a jury verdict in a court trial that says whether a defendant is guilty or not guilty rather than guilty or innocent. The fact that a person is not found guilty does not necessarily mean that he or she is innocent. Similarly, the fact that we do not reject a null hypothesis does not necessarily mean that the hypothesis is true, because another null hypothesis may be equally compatible with the data. For our P/E example, for instance, from Eq. (D.7) it is obvious any value of $\mu_X$ between 19.57 and 26.93 would be an "acceptable" hypothesis.

### A Word on Choosing the Level of Significance, $\alpha$, and the *p* Value

The Achilles heel of the classical approach to hypothesis testing is its arbitrariness in selecting $\alpha$. Although 1, 5, and 10 percent are the commonly used values of $\alpha$, there is nothing sacrosanct about these values. As noted earlier, unless we examine the consequences of committing both type I and type II errors, we cannot make the appropriate choice of $\alpha$. In practice, it is preferable to find the ***p* value** (i.e., the probability value), also known as the *exact significance level*, of the test statistic. This may be defined as *the lowest significance level at which a null hypothesis can be rejected.*

To illustrate, in an application involving 20 d.f. a $t$ value of 3.552 was obtained. The $t$ table given in Appendix E (Table E-2) shows that the $p$ value for this $t$ is 0.001 (one-tailed) or 0.002 (two-tailed). That is, this $t$ value is statistically significant at the 0.001 (one-tailed) or 0.002 (two-tailed) level.

For our P/E example under the null hypothesis that the true P/E ratio is 18.5, we found that $t = 2.43$. If the alternative hypothesis is that the true P/E ratio is greater than 18.5, we find from Table E-1 in Appendix E that $P(t > 2.43)$ is about .01 This is the $p$ value of the $t$ statistic. We say that this $t$ value is statistically significant at the 0.01 or 1 percent level. Put differently, if we were to fix $\alpha = 0.01$, at that level we can reject the null hypothesis that the true $\mu_X = 18.5$. Of course, this is a much smaller probability, smaller than the conventional $\alpha$ value, such as 5 percent. Therefore, we can reject the null hypothesis much more emphatically than if we were to choose, say, $\alpha = 0.05$. *As a rule, the smaller the p value, the stronger the evidence against the null hypothesis.*

One virtue of quoting the $p$ value is that it avoids the arbitrariness involved in fixing $\alpha$ at artificial levels, such as 1, 5, or 10 percent. If, for example, in an application the $p$ value of a test statistic (such as $t$) is, say, 0.135, and if you are willing to accept an $\alpha = 13.5$ percent, this $p$ value is statistically significant (i.e., you reject the null hypothesis at this level of significance). Nothing is wrong if you want to take a chance of being wrong 13.5 percent of the time if you reject the true null hypothesis.

Nowadays several statistical packages routinely compute the $p$ values of various test statistics, and *it is recommended that you report these p values.*

## The $\chi^2$ and $F$ Tests of Significance

Besides the $t$ test of significance discussed previously, in the main chapters of the text we will need tests of significance based on the $\chi^2$ and the $F$ probability distributions considered in Appendix C. Since the philosophy of testing is the same, we will simply present here the actual mechanism with a couple of illustrative examples; we will present further examples in the main text.

**The $\chi^2$ test of significance** In Appendix C (see Example C.14) we showed that if $S^2$ is the sample variance obtained from a random sample of $n$ observations from a normal population with variance $\sigma^2$, then the quantity

$$(n - 1)\left(\frac{S^2}{\sigma^2}\right) \sim \chi^2_{(n-1)} \tag{D.21}$$

That is, the ratio of the sample variance to population variance multiplied by the d.f. $(n - 1)$ follows the $\chi^2$ distribution with $(n - 1)$ d.f. If the d.f. and $S^2$ are known but $\sigma^2$ is not known, we can establish a $(1 - \alpha)$% confidence interval for the true but unknown $\sigma^2$ using the $\chi^2$ distribution. The mechanism is similar to that for establishing confidence intervals on the basis of the $t$ test.

But if we are given a specific value of $\sigma^2$ under $H_0$, we can directly compute the $\chi^2$ value from expression (D.21) and test its significance against the critical $\chi^2$ values given in Table E-4 in Appendix E. An example follows.

**Example D.4.**

Suppose a random sample of 31 observations from a normal population gives a (sample) variance of $S^2 = 12$. Test the hypothesis that the true variance is 9 against the hypothesis that it is different from 9. Use $\alpha = 5\%$. Here

$$H_0{:}\sigma^2 = 9 \quad \text{and} \quad H_1{:}\sigma^2 \neq 9$$

***Answer:*** Putting the appropriate numbers in expression (D.21), we obtain: $\chi^2 = 30(12/9) = 40$, which has 30 d.f. From Table E-4 in Appendix E, we observe that the probability of obtaining a $\chi^2$ value of about 40 or higher (for 30 d.f.) is 0.10 or 10 percent. Since this probability is greater than our level of significance of 5 percent, we do not reject the null hypothesis that the true variance is 9.

Table D-3 summarizes the $\chi^2$ **test** for the various types of null and alternative hypotheses.

**The *F* Test of Significance** In Appendix C we showed that if we have two randomly selected samples from two normal populations, $X$ and $Y$, with $m$ and $n$ observations, respectively, then the variable

$$F = \frac{S_X^2}{S_Y^2} = \frac{\sum(X_i - \overline{X})^2/(m-1)}{\sum(Y_i - \overline{Y})^2/(n-1)} \tag{D.22}$$

follows the $F$ distribution with $(m-1)$ and $(n-1)$ d.f., *provided the variances of the two normal populations are equal*. In other words, the $H_0$ is $\sigma_X^2 = \sigma_Y^2$. To test this hypothesis, we use the ***F* test** given in Eq. (D.22). An example follows.

**Example D.5.**

Refer to the S.A.T. math scores for male and female students given in Examples C.12 and C.15. The variances of these scores were (48.31) for the

**TABLE D-3** A SUMMARY OF THE $\chi^2$ TEST

| Null hypothesis $H_0$ | Alternative hypothesis $H_1$ | Critical region Reject $H_0$ if |
|---|---|---|
| $\sigma_X^2 = \sigma_0^2$ | $\sigma_X^2 > \sigma_0^2$ | $\dfrac{(n-1)S^2}{\sigma_0^2} > \chi^2_{\alpha,(n-1)}$ |
| $\sigma_X^2 = \sigma_0^2$ | $\sigma_X^2 < \sigma_0^2$ | $\dfrac{(n-1)S^2}{\sigma_0^2} < \chi^2_{(1-\alpha),(n-1)}$ |
| $\sigma_X^2 = \sigma_0^2$ | $\sigma_X^2 \neq \sigma_0^2$ | $\dfrac{(n-1)S^2}{\sigma_0^2} > \chi^2_{\alpha/2,(n-1)}$ or $< \chi^2_{(1-\alpha/2),(n-1)}$ |

*Note:* $\sigma_0^2$ is the value of $\sigma_X^2$ under the null hypothesis. The first subscript on $\chi^2$ in the last column is the level of significance and the second subscript is the d.f. These are critical $\chi^2$ values.

male students and (102.07) for the female students. The number of observations were 36 or 35 d.f. each. Assuming that these variances represent a sample from a much larger population of S.A.T. scores, test the hypothesis that the male and female population variances on the math part of the S.A.T. scores are the same. Use $\alpha = 1\%$.

***Answer:*** Here the $F$ value is $102.07/48.31 = 2.1128$ (approx.). This $F$ value has the $F$ distribution with 35 d.f. each. Now from Table E-3 in Appendix E we see that for 30 d.f. (35 d.f. is not given in the table), the *critical* $F$ value at the 1% level of significance is 2.39. Since the observed $F$ value of 2.1128 is less than 2.39, it is not statistically significant. That is, at $\alpha = 1\%$, we do not reject the null hypothesis that the two population variances are the same.

**Example D.6.**

In the preceding example, what is the $p$ value of obtaining an $F$ value of 2.1128? Using MINITAB, we can find that for 35 d.f. in the numerator and denominator, the probability of obtaining an $F$ value of 2.1128 or greater is about 0.01492 or about 10.5 percent. This is the $p$ value of obtaining an $F$ value of as much as 2.1128 or greater. In other words, this is the lowest level of probability at which we can reject the null hypothesis that the two variances are the same. Therefore, in this case if we reject the null hypothesis that the two variances are the same, we are taking the chance of being wrong 1.5 out of 100 times.

Examples D.5 and D.6 suggest a practical strategy. We may fix $\alpha$ at some level (e.g., 1, 5, or 10 percent) and also find out the $p$ value of the test statistic. If the estimated $p$ value is smaller than the chosen level of significance, we can reject the null hypothesis under consideration. On the other hand, if the estimated $p$ value is greater than the preselected level of significance, we may not reject the null hypothesis.

Table D-4 summarizes the $F$ test.

**TABLE D-4** A SUMMARY OF THE $F$ STATISTIC

| Null hypothesis $H_0$ | Alternative hypothesis $H_1$ | Critical region Reject $H_0$ if |
|---|---|---|
| $\sigma_1^2 = \sigma_2^2$ | $\sigma_1^2 > \sigma_2^2$ | $\frac{S_1^2}{S_2^2} > F_{\alpha,n\text{df},\,d\text{df}}$ |
| $\sigma_1^2 = \sigma_2^2$ | $\sigma_1^2 \neq \sigma_2^2$ | $\frac{S_1^2}{S_2^2} > F_{\alpha/2,n\text{df},d\text{df}}$ or $< F_{(1-\alpha/2),n\text{df},d\text{df}}$ |

*Notes:*

1. $\sigma_1^2$ and $\sigma_2^2$ are the two population variances.
2. $S_1^2$ and $S_2^2$ are the two sample variances.
3. $n$df and $d$df denote, respectively, the numerator and denominator d.f.
4. In computing the $F$ ratio, put the larger $S^2$ value in the numerator.
5. The critical $F$ values are given in the last column. The first subscript of $F$ is the level of significance and the second subscript is the numerator and denominator d.f.
6. Note that $F_{(1-\alpha/2),n\text{df},d\text{df}} = \frac{1}{F_{\alpha/2,d\text{df},n\text{df}}}$.

To conclude this appendix, we summarize the steps involved in testing a statistical hypothesis:

**Step 1:** State the null hypothesis $H_0$ and the alternative hypothesis $H_1$ (e.g., $H_0$:$\mu_X = 18.5$ and $H_1$:$\mu_X \neq 18.5$ for our P/E example).
**Step 2:** Select the test statistic (e.g., $\overline{X}$).
**Step 3:** Determine the probability distribution of the test statistic (e.g., $\overline{X} \sim N(\mu_X, \sigma^2_{\overline{X}}/n)$.
**Step 4:** Choose the level of significance $\alpha$, that is, the probability of committing a type I error. (But keep in mind our discussion about the *p* value.)
**Step 5:** Choose the confidence interval or the test of significance approach.

**The Confidence Interval Approach** Using the probability distribution of the test statistic, establish a $100(1 - \alpha)\%$ confidence interval. If this interval (i.e., the acceptance region) includes the *null-hypothesized value,* do not reject the null hypothesis. But if this interval does not include it, reject the null hypothesis.

**The Test of Significance Approach** Alternatively, you can follow this approach by obtaining the relevant test statistic (e.g., the *t* statistic) under the null hypothesis and find out the *p* value of obtaining a specified value of the test statistic from the appropriate probability distribution (e.g., the *t, F,* or the $\chi^2$ distribution). If this probability is less than the prechosen value of $\alpha$, you can reject the null hypothesis. But if it is greater than $\alpha$, do not reject it. If you do not want to preselect $\alpha$, just present the *p* value of the statistic.

Whether you choose the confidence interval or the test of significance approach, *always keep in mind that in rejecting or not rejecting a null hypothesis you are taking a chance of being wrong* $\alpha$ (*or p value) percent of the time.*

Further uses of the various tests of significance discussed in this appendix will be illustrated throughout the rest of this book.

## D.6 SUMMARY

Estimating population parameters on the basis of sample information and testing hypotheses about them in light of the sample information are the two main branches of (classical) statistical inference. In this appendix we examined the essential features of these branches.

## KEY TERMS AND CONCEPTS

The key terms and concepts introduced in this appendix are

Statistical inference
Parameter estimation
  **a)** point estimation
  **b)** interval estimation
Sampling (probability) distribution
Critical *t* values
Confidence interval (CI)
  **a)** confidence coefficient
  **b)** random interval (lower limit, upper limit)

Level of significance
Probability of committing a type I error
Properties of estimators
- **a)** linearity (linear estimator)
- **b)** unbiasedness (unbiased estimator)
- **c)** minimum variance (minimum-variance estimator)
- **d)** efficiency (efficient estimator)
- **e)** best linear unbiased estimator (BLUE)
- **f)** consistency (consistent estimator)

Hypothesis testing
- **a)** null hypothesis
- **b)** alternative hypothesis
- **c)** one-sided; one-tailed hypothesis
- **d)** two-sided; two-tailed; composite hypothesis

Confidence interval (approach to hypothesis testing)
- **a)** acceptance region
- **b)** critical region; region of rejection
- **c)** critical values

Type I error ($\alpha$); level of significance; confidence coefficient ($1 - \alpha$)
Type II error ($\beta$)
power of the test ($1 - \beta$)
Tests of significance (approach to hypothesis testing)
- **a)** Test statistic; $t$ statistic; $t$ test
- **b)** $\chi^2$ test
- **c)** $F$ test

The $p$ value

## QUESTIONS

**D.1.** What is the distinction between each of the following pairs of terms?
- **a.** Point estimator and interval estimator.
- **b.** Null and alternative hypotheses.
- **c.** Type I and type II errors.
- **d.** Confidence coefficient and level of significance.
- **e.** Type II error and power.

**D.2.** What is the meaning of
- **a.** Statistical inference.
- **b.** Sampling distribution.
- **c.** Acceptance region.
- **d.** Test statistic.
- **e.** Critical value of a test.
- **f.** Level of significance.
- **g.** The $p$ value.

**D.3.** Explain carefully the meaning of
- **a.** An unbiased estimator.
- **b.** A minimum variance estimator.
- **c.** A best, or efficient, estimator.
- **d.** A linear estimator.
- **e.** A best linear unbiased estimator (BLUE).

**D.4.** State whether the following statements are true, false, or uncertain. Justify your answers.
- **a.** An estimator of a parameter is a random variable, but the parameter is non-random, or fixed.
- **b.** An unbiased estimator of a parameter, say, $\mu_X$, means that it will always be equal to $\mu_X$.
- **c.** An estimator can be a minimum variance estimator without being unbiased.
- **d.** An efficient estimator means an estimator with minimum variance.
- **e.** An estimator can be BLUE only if its sampling distribution is normal.
- **f.** An acceptance region and a confidence interval for any given problem means the same thing.

**g.** A type I error occurs when we reject the null hypothesis even though it is false.
**h.** A type II error occurs when we reject the null hypothesis even though it may be true.
**i.** As the degrees of freedom (d.f.) increase indefinitely, the $t$ distribution approaches the normal distribution.
**j.** The central limit theorem states that the sample mean is always distributed normally.
**k.** The terms *level of significance* and *p value* mean the same thing.

**D.5.** Explain carefully the difference between the confidence interval and test of significance approaches to hypothesis testing.

**D.6.** Suppose in an example with 40 d.f. that you obtained a $t$ value of 1.35. Since its $p$ value is somewhere between a 5 and 10 percent level of significance (one-tailed), it is not statistically very significant. Do you agree with this statement? Why or why not?

## PROBLEMS

**D.7.** Find the critical $Z$ values in the following cases:
**a.** $\alpha = 0.05$ (two-tailed test)
**b.** $\alpha = 0.05$ (one-tailed test)
**c.** $\alpha = 0.01$ (two-tailed test)
**d.** $\alpha = 0.02$ (one-tailed test)

**D.8.** Find the critical $t$ values in the following cases:
**a.** $n = 4, \alpha = 0.05$ (two-tailed test)
**b.** $n = 4, \alpha = 0.05$ (one-tailed test)
**c.** $n = 14, \alpha = 0.01$ (two-tailed test)
**d.** $n = 14, \alpha = 0.01$ (one-tailed test)
**e.** $n = 60, \alpha = 0.05$ (two-tailed test)
**f.** $n = 200, \alpha = 0.05$ (two-tailed test)

**D.9.** Assume that the per capita income of residents in a country is normally distributed with mean $\mu = \$1000$ and variance $\sigma^2 = 10{,}000$ ($ squared).
**a.** What is the probability that the per capita income lies between $800 and $1200?
**b.** What is the probability that it exceeds $1200?
**c.** What is the probability that it is less than $800?
**d.** Is it true that the probability of per capita income exceeding $5000 is practically zero?

**D.10.** Continuing with problem D.9, based on a random sample of 1000 members, suppose that you find the sample mean income, $\overline{X}$, to be $900.
**a.** Given that $\mu = \$1000$, what is the probability of obtaining such a sample mean value?
**b.** Based on the sample mean, establish a 95% confidence interval for $\mu$ and find out if this confidence interval includes $\mu = \$1000$. If it does not, what conclusions would you draw?
**c.** Using the test of significance approach, decide whether you want to accept or reject the hypothesis that $\mu = \$1000$. Which test did you use and why?

**D.11.** The number of peanuts contained in a jar follows the normal distribution with mean $\mu$ and variance $\sigma^2$. Quality control inspections over several periods show that 5 percent of the jars contain less than 6.5 ounces of peanuts and 10 percent contain more than 6.8 ounces.
**a.** Find $\mu$ and $\sigma^2$.
**b.** What percentage of bottles contain more than 7 ounces?

**D.12.** The following random sample was obtained from a normal population with mean $\mu$ and variance = 2.

8, 9, 6, 13, 11, 8, 12, 5, 4, 14

**a.** Test: $\mu = 5$ against $\mu \neq 5$
**b.** Test: $\mu = 5$ against $\mu > 5$
*Note:* use $\alpha = 5\%$.
**c.** What is the $p$ value in part (*a*) of this problem?

**D.13.** Based on a random sample of 10 values from a normal population with mean $\mu$ and standard deviation $\sigma$, you calculated that $\overline{X} = 8$ and the sample standard deviation = 4. Estimate a 95% confidence interval for the population mean. Which probability distribution did you use? Why?

**D.14.** You are told that $X \sim N(\mu_X = 8, \sigma_X^2 = 36)$. Based on a sample of 25 observations, you found that $\overline{X} = 7.5$.
**a.** What is the sampling distribution of $\overline{X}$?
**b.** What is the probability of obtaining an $\overline{X} = 7.5$ or less?
**c.** From your answer in part (*b*) of this problem, could such a sample value have come from the preceding population?

**D.15.** Compute the $p$ values in the following cases:
**a.** $t \geq 1.72$, d.f. = 24
**b.** $Z \geq 2.9$
**c.** $F \geq 2.59$, d.f. = 3 and 20, respectively
**d.** $\chi^2 \geq 19$, d.f. = 30
*Note:* If you cannot get an exact answer from the various probability distribution tables, try to obtain them from a program such as MINITAB or Excel.

**D.16.** In an application involving 30 d.f. you obtained a $t$ statistic of 0.68. Since this $t$ value is not statistically significant even at the 10% level of significance, you can safely accept the relevant hypothesis. Do you agree with this statement? What is the $p$ value of obtaining such a statistic?

**D.17.** Let $X \sim N(\mu_X, \sigma_X^2)$. A random sample of three observations was obtained from this population. Consider the following estimators of $\mu_X$:

$$\hat{\mu}_1 = \frac{X_1 + X_2 + X_3}{3} \quad \text{and} \quad \hat{\mu}_2 = \frac{X_1}{6} + \frac{X_2}{3} + \frac{X_3}{2}$$

**a.** Is $\hat{\mu}_1$ an unbiased estimator of $\mu_X$? What about $\hat{\mu}_2$?
**b.** If both estimators are unbiased, which one would you choose? (*Hint:* Compare the variances of the two estimators.)

**D.18.** Refer to Problem C.10 in Appendix C. Suppose a random sample of 10 firms gave a mean profit of $900,000 and a (sample) standard deviation of $100,000.
**a.** Establish a 95% confidence interval for the true mean profit in the industry.
**b.** Which probability distribution did you use? Why?

**D.19.** Refer to Example C.14 in Appendix C.
**a.** Establish a 95% confidence interval for the true $\sigma^2$.
**b.** Test the hypothesis that the true variance is 8.2.

**D.20.** Sixteen cars are first driven with a standard fuel and then with Petrocoal, a gasoline with a methanol additive. The results of the nitrous oxide emissions ($NO_x$) test are as follows:

| Type of fuel | Average $NO_x$ | Standard deviation of $NO_x$ |
|---|---|---|
| Standard | 1.075 | 0.5796 |
| Petrocoal | 1.159 | 0.6134 |

*Source:* Michael O. Finkelstein and Bruce Levin, *Statistics for Lawyers,* Springer-Verlag, New York, 1990, p. 230.

**a.** How would you test the hypothesis that the two population standard deviations are the same?

**b.** Which test did you use? What are the assumptions underlying that test?

**D.21.** Show that the estimator given in Eq. (D.14) is biased. (*Hint:* Expand Eq. (D.14), and take the expectation of each term, keeping in mind that the expected value of each $X_i$ is $\mu_X$).

**D.22.** *One-sided confidence interval*. Return to the P/E example in this appendix and look at the two-sided 95% confidence interval given in Eq. (D.7). Suppose you want to establish a one-sided confidence interval only, either an upper bound or a lower bound. How would you go about establishing such an interval? (*Hint:* Find the one-tail critical $t$ value.)