APPENDIX C

SOME IMPORTANT PROBABILITY DISTRIBUTIONS

In Appendix B we noted that a random variable (r.v.) can be described by a few characteristics, or moments, of its probability function (PDF or PMF), such as the expected value and variance. This, however, presumes that we know the PDF of that r.v., which is a tall order since there are all kinds of random variables. In practice, however, some random variables occur so frequently that statisticians have determined their PDFs and documented their properties. For our purpose, we will consider only those PDFs that are of direct interest to us. But keep in mind that there are several other PDFs that statisticians have studied which can be found in any standard statistics textbook. In this appendix we will discuss the following four probability distributions:

1. The normal distribution
2. The t distribution
3. The chi-square (χ^2) distribution
4. The F distribution

These probability distributions are important in their own right, but for our purposes they are especially important because they help us to find out the probability distributions of estimators (or *statistics*), such as the sample mean and sample variance. Recall that estimators are random variables. Equipped with that knowledge, we will be able to draw inferences about their true population values. For example, if we know the probability distribution of the sample mean, $\overline{X}$, we will be able to draw inferences about the true, or population, mean μ_X. Similarly, if we know the probability distribution of the sample variance S_x^2, we will be able to say something about the true population variance, σ_X^2. This is the essence of *statistical inference,* or drawing conclusions about some characteristics (i.e., moments) of the population on the basis of the sample at hand. We will discuss in depth how this is accomplished in Appendix D. For now we discuss the salient features of the four probability distributions.

C.1 THE NORMAL DISTRIBUTION

Perhaps the single most important probability distribution involving a continuous r.v. is the **normal distribution.** Its *bell-shaped* picture, as shown in Figure A-3, should be familiar to anyone with a modicum of statistical knowledge. Experience has shown that the normal distribution is a reasonably good model for a continuous r.v. whose value depends on a number of factors, each factor exerting a comparatively small positive or negative influence. Thus, consider the r.v. body weight. It is likely to be normally distributed because factors such as heredity, bone structure, diet, exercise, and metabolism are each expected to have some influence on weight, yet no single factor dominates the others. Likewise, variables such as height and grade-point average are also found to be normally distributed.

For notational convenience, we express a normally distributed r.v. X as

$$X \sim N\left(\mu_X, \sigma_X^2\right) \qquad \textbf{(C.1)}^{1}$$

where $\sim$ means distributed as, N stands for the normal distribution, and the quantities inside the parentheses are the *parameters* of the distribution, namely, its (population) mean or expected value μ_X and its variance σ_X^2. Note that X is a continuous r.v. and may take any value in the range $-\infty$ to ∞.

Properties of the Normal Distribution

1. The normal distribution curve, as Figure A-3 shows, is symmetrical around its mean value μ_X.
2. The PDF of a normally distributed r.v. is highest at its mean value but tails off at its extremities (i.e., in the tails of the distribution). That is, the probability of obtaining a value of a normally distributed r.v. far away from its mean value becomes progressively smaller. For example, the probability of someone exceeding the height of 7.5 feet is very small.
3. As a matter of fact, *approximately* 68 percent of the area under the normal curve lies between the values of $(\mu_X \pm \sigma_X)$, approximately 95 percent of the area lies between $(\mu_X \pm 2\sigma_X)$, and approximately 99.7 percent of the area lies between $(\mu_X \pm 3\sigma_X)$, as shown in Figure C-1. As noted in Appendix A, and

[1]For the mathematically inclined student, here is the mathematical equation for the PDF of a normally distributed r.v. X:

$$f(X) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{X-\mu_X}{\sigma_X}\right)^2\right\}$$

where exp{ } means e raised to the power the expression inside { }, $e \approx 2.71828$ (the base of natural logarithm), and $\pi \approx 3.14159$. μ_X and σ_X^2, known as the parameters of the distribution, are, respectively, the mean, or expected value, and the variance of the distribution.

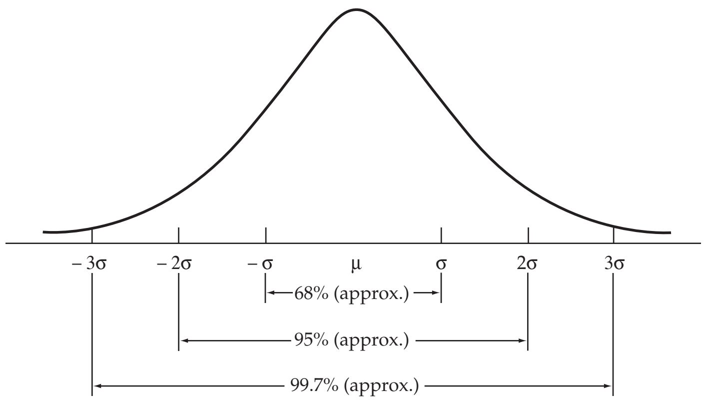

FIGURE C-1 Areas under the normal curve

discussed further subsequently, these areas can be used as measures of probabilities. The total area under the curve is 1, or 100 percent.

4. A normal distribution is *fully described by its two parameters,* μ_X and σ_X^2. That is, once the values of these two parameters are known, we can find out the probability of X lying within a certain interval from the mathematical formula given in footnote 1. Fortunately, we do not have to compute the probabilities from this formula because these probabilities can be obtained easily from the specially prepared table in Appendix E (Table E-1). We will explain how to use this table shortly.
5. A *linear combination (function) of two (or more) normally distributed random variables is itself normally distributed*—an especially important property of the normal distribution in econometrics. To illustrate, let

$$X \sim N\left(\mu_X, \sigma_X^2\right)$$
$$Y \sim N\left(\mu_Y, \sigma_Y^2\right)$$

and assume that X and Y are *independent.*[2]

Now consider the linear combination of these two variables: $W = aX + bY$, where a and b are constant (e.g., $W = 2X + 4Y$); then

$$W \sim N[\mu_W, \sigma_W^2] \tag{C.2}$$

where

$$\mu_W = (a\mu_X + b\mu_Y)$$
$$\sigma_W^2 = \left(a^2\sigma_X^2 + b^2\sigma_Y^2\right) \tag{C.3}$$

[2]Recall that two variables are independently distributed if their joint PDF (PMF) is the product of their marginal PDFs, that is, $f(X, Y) = f(X)f(Y)$, for all values of X and Y.

Note that in Eq. (C.3) we have used some of the properties of the expectation operator E and the variances of independent random variables discussed in Appendix B. (See Section B.2.)[3] Incidentally, expression (C.2) can be extended straightforwardly to a linear combination of more than two normal random variables.

6. For a normal distribution, skewness (S) is zero and kurtosis (K) is 3.

Example C.1.

Let X denote the number of roses sold daily by a florist in uptown Manhattan and Y the number of roses sold daily by a florist in downtown Manhattan. Assume that both X and Y are *independently* normally distributed as $X \sim N(100, 64)$ and $Y \sim N(150, 81)$. What is the average value of the roses sold in two days by the two florists and the corresponding variance of sale? Here $W = 2X + 2Y$. Therefore, following expression (C.3), we have $E(W) = E(2X + 2Y) = 500$ and $\text{var}(W) = 4\ \text{var}(X) + 4\ \text{var}(Y) = 580$. Therefore, W is distributed normally with a mean value of 500 and a variance of 580: $[W \sim N(500, 580)]$.

The Standard Normal Distribution

Although a normal distribution is fully specified by its two parameters, (population) mean or expected value and variance, one normal distribution can differ from another in either its mean or variance, or both, as shown in Figure C-2.

How do we compare the various normal distributions shown in Figure C-2? Since these normal distributions differ in either their mean values or variances, or both, let us define a new variable, Z, as follows:

$$Z = \frac{X - \mu_X}{\sigma_X} \tag{C.4}$$

If the variable X has a mean μ_X and a variance σ_X^2, it can be shown that the Z variable defined previously has a mean value of zero and a variance of 1 (or unity). (For proof, see Problem C.26). In statistics such a variable is known as a **unit** or **standardized variable.**

If $X \sim N(\mu_X, \sigma_X^2)$, then Z as defined in Eq. (C.4) is known as a **unit** or **standard normal variable**, that is, a normal variable with zero mean and unit (or 1) variance. We write such a normal variable as:

$$Z \sim N(0, 1) \qquad \textbf{(C.5)}^4$$

[3]Note that if X and Y are normally distributed but are not independent, W is still normally distributed with the mean given in Eq. (C.3) but with the following variance (cf. Eq. B.27): $\sigma_w^2 = a^2\sigma_X^2 + b^2\sigma_y^2 + 2ab\ \text{cov}\,(X, Y)$.

[4]This can be proved easily by noting the property of the normal distribution that a linear function of a normally distributed variable is itself normally distributed. Note that given μ_X and σ_X^2, Z is a linear function of X.

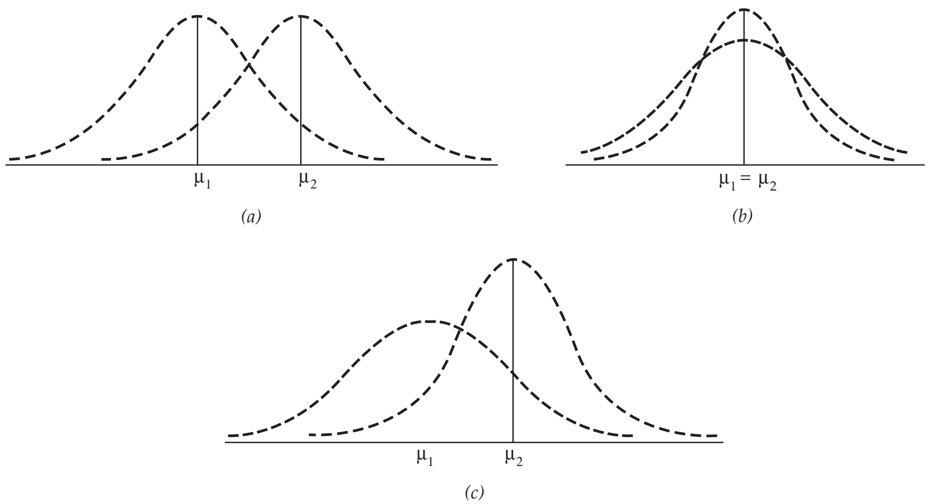

FIGURE C-2 (*a*) Different means, same variance; (*b*) same mean, different variances; (*c*) different means, different variances

Thus, *any normally distributed r.v. with a given mean and variance can be converted to a standard normal variable,* which greatly simplifies our task of computing probabilities, as we will show shortly.

The PDF and CDF (cumulative distribution function) of the standard normal distribution are shown in Figures C-3(*a*) and C-3(*b*), respectively. (See Section A.5 on the definitions of PDF and CDF. See also Tables E-1(*a*) and E-1(*b*) in Appendix E.) The CDF, like any other CDF, gives the probability that the standard normal variable takes a value equal to or less than z, that is, $P(Z \leq z)$, where z is a specific numerical value of Z.

To illustrate how we use the standard normal distribution to compute various probabilities, we consider several concrete examples.

Example C.2.

It is given that X, the daily sale of bread in a bakery, follows the normal distribution with a mean of 70 loaves and a variance of 9; that is, $X \sim N(70, 9)$. What is the probability that on any given day the sale of bread is greater than 75 loaves?

Since X follows the normal distribution with the stated mean and variance, it follows that

$$Z = \frac{75 - 70}{3} = \approx 1.67$$

follows the standard normal distribution. Therefore, we want to find[5]

$$P(Z > 1.67)$$

[5]*Note:* Whether we write $P(Z > 1.67)$ or $P(Z \geq 1.67)$ is immaterial because, as noted in Appendix A, the probability that a continuous r.v. takes a particular value (e.g., 1.67) is always zero.

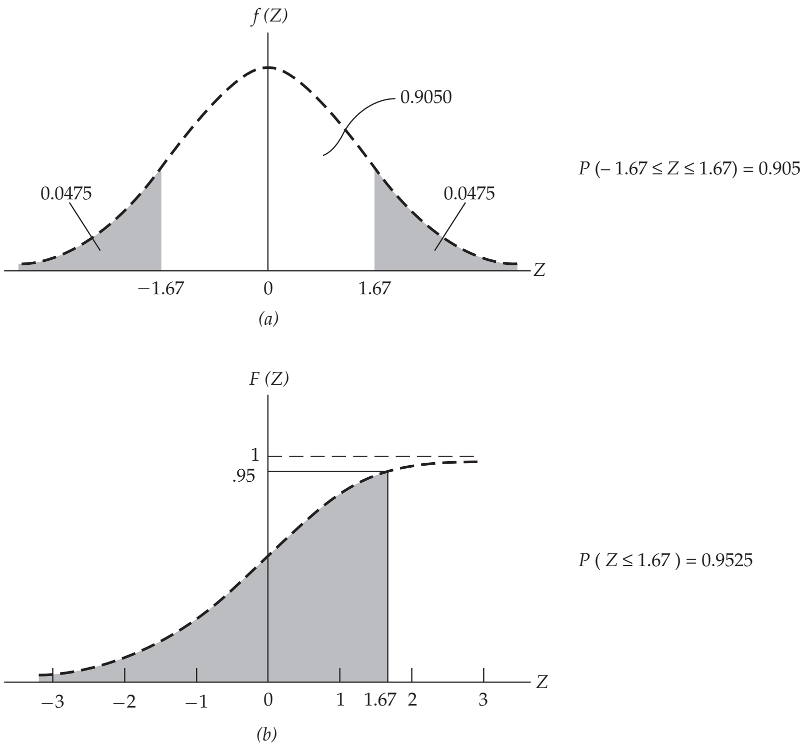

FIGURE C-3 (*a*) PDF and (*b*) CDF of the standard normal variable

Now Table E-1(*b*) in Appendix E gives the CDF of the standard normal distribution between the values of $Z = -3.0$ to $Z = 3.0$. For example, this table shows that the probability that Z lies between -3.0 to $Z = 1.67 = 0.9525$. Therefore,

$$P(Z > 1.67) = 1 - 0.9525 = 0.0475$$

That is, the probability of the daily sale of bread exceeding 75 loaves of bread is 0.0475 or about 4.75 percent. (See Figure C-3[*a*].)

Example C.3.

Continue with Example C.2, but suppose we now want to find out the probability of a daily sale of bread of 75 or fewer loaves. The answer is obvious from the previous example, namely, that this probability is 0.9525 which is shown in Figure C-3(*b*).

Example C.4.

Continue with Example C.2, but now suppose we want to find out the probability that the daily sale of bread is between 65 and 75 loaves. To compute this probability, we first compute

$$Z_1 = \frac{65 - 70}{3} = \approx -1.67$$

$$Z_2 = \frac{75 - 70}{3} = \approx 1.67$$

Now from Table E-1 we see that

$$P(-3.0 \le Z \le -1.67) = 0.0475$$

and

$$P(-3.0 \le Z \le 1.67) = 0.9525$$

Therefore,

$$P(-1.67 \le Z \le 1.67) = 0.9525 - 0.0475 = 0.9050$$

That is, the probability is 90.5 percent that the sales volume will lie between 65 and 75 loaves of bread per day, as shown in Figure C-3(*a*).

Example C.5.

Continue with the preceding example but now assume that we want to find the probability that the sale of bread either exceeds 75 loaves or is less than 65 loaves per day. If you have mastered the previous examples, you can see easily that this probability is 0.0950, as shown in Figure C-3(*a*).

As the preceding examples show, once we know that a particular r.v. follows the normal distribution with a given mean and variance, all we have to do is convert that variable into the standard normal variable and compute the relevant probabilities from the standard normal table (Table E-1). It is indeed remarkable that just one standard normal distribution table suffices to deal with any normally distributed variable regardless of its specific mean and variance values.

As we have remarked earlier, the normal distribution is probably the single most important theoretical probability distribution because several (continuous) random variables are found to be normally distributed or at least approximately so. We will show this in Section C.2. But before that, we consider some practical problems in dealing with the normal distribution.

TABLE C-1 25 RANDOM NUMBERS FROM $N(0, 1)$ AND $N(2, 4)$

$N(0, 1)$	$N(2, 4)$	$N(0, 1)$	$N(2, 4)$
–0.48524	4.25181	0.22968	0.21487
0.46262	0.01395	–0.00719	–0.47726
2.23092	0.09037	–0.71217	1.32007
–0.23644	1.96909	–0.53126	–1.25406
1.10679	1.62206	–1.02664	3.09222
–0.82070	1.17653	–1.29535	1.05375
0.86553	2.78722	–0.61502	0.58124
–0.40199	2.41138	–1.80753	1.55853
1.13667	2.58235	0.20687	1.71083
–2.05585	0.40786	–0.19653	0.90193
2.98962	0.24596	2.49463	–0.14726
0.61674	–3.45379	0.94602	–3.69238
–0.32833	3.29003		

Random Sampling from a Normal Population

Since the normal distribution is used so extensively in theoretical and practical statistics, it is important to know how we can obtain a random sample from such a population. Suppose we wish to draw a random sample of 25 observations from a normal probability distribution with a mean of zero and variance of 1 [i.e., the standard normal distribution, $N(0, 1)$]. How do we obtain such a sample?

Most statistical packages have routines, called **random number generators,** to obtain random samples from the most frequently used probability distributions. For example, using the MINITAB statistical package, we obtained 25 random numbers from an $N(0, 1)$ normal population. These are shown in the first column of Table C-1. Also shown in column 2 of the table is another random sample of 25 observations obtained from a normal population with mean 2 and variance 4 (i.e., $N(2, 4)$).[6] Of course, you can generate as many samples as wanted by the procedure just described.

The Sampling or Probability Distribution of the Sample Mean $\bar{X}$

In Appendix B we introduced the sample mean (see Eq. [B.43]) as an estimator of the population mean. But since the sample mean is based on a given sample, its value will vary from sample to sample; that is, the sample mean can be treated as an r.v., which will have its own PDF. Can we find out the PDF of the sample mean? The answer is yes, provided the sample is drawn randomly.

[6]MINITAB will generate a random sample from a normal population with a given mean variance. Actually, once we obtain a random sample from the standard normal distribution [i.e., $N(0, 1)$], we can easily convert this sample to a normal population with a different mean and variance. Let $Y = a + bZ$, where Z is $N(0, 1)$, and where a and b are constants. Since Y is a linear combination of a normally distributed variable, Y is itself normally distributed with $E(Y) = E(a + bZ) = a$, since $E(Z) = 0$ and var $(a + bZ) = b^2$ var $(Z) = b^2$, since var $(Z) = 1$. Hence, $Y \sim N(a, b^2)$. Therefore, if you multiply the values of Z by b and add a to it, you will have a sample from a normal population with mean a and variance b^2. Thus, if $a = 2$ and $b = 2$, we have $Y \sim N(2, 4)$.

In Appendix B we described the notion of random sampling in an intuitive way by letting each member of the population have an equal chance of being included in the sample. In statistics, however, the term **random sampling** is used in a rather special sense. *We say that $X_1, X_2, \ldots, X_n$ constitutes a random sample of size n if all these X's are drawn independently from the same probability distribution (i.e., each X_i has the same PDF).* The X's thus drawn are known as **i.i.d.** (*independently and identically distributed*) **random variables.** In the remainder of this appendix and the main chapters of the text, therefore, the term *random sample* will denote a sample of i.i.d. random variables. For brevity, sometimes we will use the term an *i.i.d. sample* to mean a random sample in the sense just described.

Thus, if each $X_i \sim N(\mu_X, \sigma_X^2)$ and if each X_i value is drawn independently, then we say that $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, the normal PDF being their common probability distribution. Note two things about this definition: First, each X included in the sample must have the *same* PDF and, second, each X included in the sample is drawn independently of the others.

Given the very important concept of random sampling, we now develop another very important concept in statistics, namely, the concept of the **sampling,** or **probability, distribution of an estimator,** such as, say, the sample mean, $\bar{X}$. A firm comprehension of this concept is absolutely essential to understand the topic of statistical inference in Appendix D and for our discussion of econometrics in the main chapters of the text. Since many students find the concept of sampling distribution somewhat bewildering, we will explain it with an example.

Example C.6.

Consider a normal distribution with a mean value of 10 and a variance of 4, that is, $N(10, 4)$. From this population we obtain 20 random samples of 20 observations each. For each sample thus drawn, we obtain the sample mean value, $\bar{X}$. Thus we have a total of 20 sample means. These are collected in Table C-2.

Let us group these 20 means in a frequency distribution, as shown in Table C-3.

The frequency distribution of the sample means given in Table C-3 may be called the *empirical sampling,* or *probability, distribution* of the sample means.[7] Plotting this empirical distribution, we obtain the bar diagram shown in Figure C-4.

If we connect the heights of the various bars shown in the figure, we obtain the *frequency polygon,* which *resembles* the shape of the normal distribution. If we had drawn many more such samples, would the frequency polygon take the familiar bell-shaped curve of the normal distribution? That is, would the sampling distribution of the sample mean in fact follow the normal distribution? Indeed, this is the case.

[7]The sampling distribution of an estimator is like the probability distribution of any random variable, except that the random variable in this case happens to be an estimator or a statistic. Put differently, a *sampling distribution is a probability distribution where the random variable is an estimator,* such as the sample mean or sample variance.

TABLE C-2 20 SAMPLE MEANS FROM $N(10, 4)$

Sample means ($\bar{X}_i$)	
9.641	10.134
10.040	10.249
9.174	10.321
10.840	10.399
10.480	9.404
11.386	8.621
9.740	9.739
9.937	10.184
10.250	9.765
10.334	10.410

Sum of 20 sample means = 201.05

$$\bar{\bar{X}} = \frac{201.05}{20} = 10.052$$

$$\text{Var}(\bar{X}_i) = \frac{\Sigma(\bar{X}_i - \bar{\bar{X}})^2}{19}$$

$$= 0.339 \quad \textit{Note}: \bar{\bar{X}} = \frac{\Sigma \bar{X}_i}{n}$$

TABLE C-3 FREQUENCY DISTRIBUTION OF 20 SAMPLE MEANS

Range of sample mean	Absolute frequency	Relative frequency
8.5–8.9	1	0.05
9.0–9.4	1	0.05
9.5–9.9	5	0.25
10.0–10.4	8	0.40
10.5–10.9	4	0.20
11.0–11.4	1	0.05
Total	20	1.00

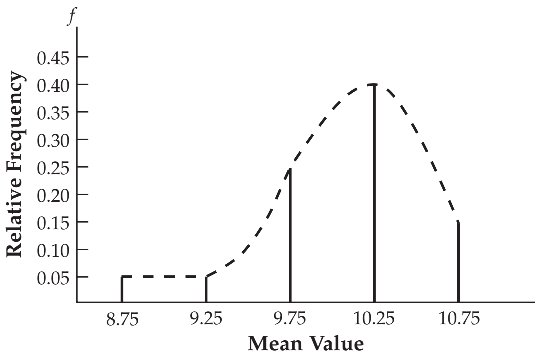

FIGURE C-4 Distribution of 20 sample means from $N(10, 4)$ population

Here we rely on statistical theory: If $X_1, X_2, \ldots, X_n$ is a *random sample* from a normal population with mean μ_X and variance σ_X^2, then the sample mean, $\bar{X}$, also follows the normal distribution with the same mean μ_X but with variance $\frac{\sigma_X^2}{n}$, that is,

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right) \tag{C.6}$$

In other words, the sampling (or probability) distribution of the sample mean, $\bar{X}$ the estimator of μ_X, also follows the normal distribution with the same mean as that of each X_i but with variance equal to the variance of $X_i (= \sigma_X^2)$ *divided by* the sample size n (for proof, see Problem C.25). As you can see for $n > 1$, the variance of the sample mean will be much smaller than the variance of any X_i. To see this graphically, go to **www.ruf.rice.edu/~lane/stat_sim** and ruin the sampling distribution applet. This will demonstrate how the distribution of sample means changes for different population distributions and different sample sizes.

If we take the (positive) square root of the variance of $\overline{X}$, we obtain $\frac{\sigma_X}{\sqrt{n}}$, which is called the **standard error** (se) of $\overline{X}$, which is akin to the concept of standard deviation. Historically, the square root of the variance of a random variable is called the standard deviation and the square root of the variance of an estimator is called the standard error. Since an estimator is also a random variable, there is no need to distinguish the two terms. But we will keep the distinction because it is so well entrenched in statistics.

Returning to our example, then, the expected value of $\overline{X}_i$, $E(\overline{X}_i)$, should be 10, and its variance should be $4/20 = 0.20$. If we take the mean value of the 20 sample means given in Table C-2, call it the *grand mean* $\overline{\overline{X}}$, it should be about equal to $E(\overline{X}_i)$, and if we compute the sample variance of these 20 sample means, it should be about equal to 0.20. As Table C-2 shows, $\overline{\overline{X}} = 10.052$, about equal to the expected value of 10 and $\mathrm{var}(\overline{X}_i) = 0.339$, which is not quite close to 0.20. Why the difference?

Notice that the data given in Table C-2 is based only on 20 samples. As noted, if we had many more samples (each based on 20 observations), we would come close to the theoretical result of mean 10 and variance of 0.20. It is comforting to know that we have such a useful theoretical result. As a consequence, we do not have to conduct the type of sampling experiment shown in Table C-2, which can be time-consuming. Just based on one random sample from the normal distribution, we can say that the expected value of the sample mean is equal to the true mean value of μ_X. As we will show in Appendix D, knowledge that a particular estimator follows a particular probability distribution will immensely help us in relating a sample quantity to its population counterpart. In passing, note that as a result of Eq. (C.6), it follows at once that

$$Z = \frac{(\overline{X} - \mu_X)}{\frac{\sigma_X}{\sqrt{n}}} \sim N(0, 1) \qquad \textbf{(C.7)}$$

that is, a standard normal variable. Therefore, you can easily compute from the standard normal distribution table the probabilities that a given sample mean is greater than or less than a given population mean. An example follows.

Example C.7.

Let X denote the number of miles per gallon achieved by cars of a particular model. You are told that $X \sim N(20, 4)$. What is the probability that, for a random sample of 25 cars, the average gallons per mile will be

a. greater than 21 miles
b. less than 18 miles
c. between 19 and 21 miles?

Since X follows the normal distribution with mean = 20 and variance = 4, we know that $\overline{X}$ also follows the normal distribution with mean = 20 and variance = 4/25. As a result, we know that

$$Z = \frac{\overline{X} - 20}{\sqrt{4/25}} = \frac{\overline{X} - 20}{0.4} \sim N(0, 1)$$

That is, Z follows the standard normal distribution. Therefore, we want to find

$$\begin{aligned} P(\overline{X} > 21) &= P\left(Z > \frac{21 - 20}{0.4}\right) \\ &= P(Z > 2.5) \\ &= 0.062 \text{ (From Table E =1[}b\text{])} \\ P(\overline{X} < 18) &= P\left(Z < \frac{18 - 20}{0.4}\right) \\ &= P(Z < -5) \approx 0 \\ P(19 \le \overline{X} \le 21) &= P(-2.5 \le Z \le 2.5) \\ &= 0.9876 \end{aligned}$$

Before moving on, note that the sampling experiment we conducted in Table C-2 is an illustration of the so-called **Monte Carlo experiments** or **Monte Carlo simulations**. They are a very inexpensive method of studying properties of various statistical models, especially when conducting real experiments would be time-consuming and expensive (see Problems C.21, C.22, and C.23).

The Central Limit Theorem (CLT)

We have just shown that the sample mean of a sample drawn from a normal population also follows the normal distribution. But what about samples drawn from other populations? There is a remarkable theorem in statistics—the **central limit theorem (CLT)**—originally proposed by the French mathematician Laplace, which states that if $X_1, X_2, \ldots, X_n$ is a random sample from *any* population (i.e., probability distribution) with mean μ_X and σ_X^2, the sample mean $\overline{X}$ tends to be normally distributed with mean μ_X and variance $\frac{\sigma_X^2}{n}$ *as the sample size increases indefinitely* (technically, infinitely).[8] Of course, if the X_i happen to be from the normal population, the sample mean follows the normal distribution regardless of the sample size. This is shown in Figure C-5.

[8]In practice, no matter what the underlying probability distribution is, the sample mean of a sample size of at least 30 observations will be approximately normal.

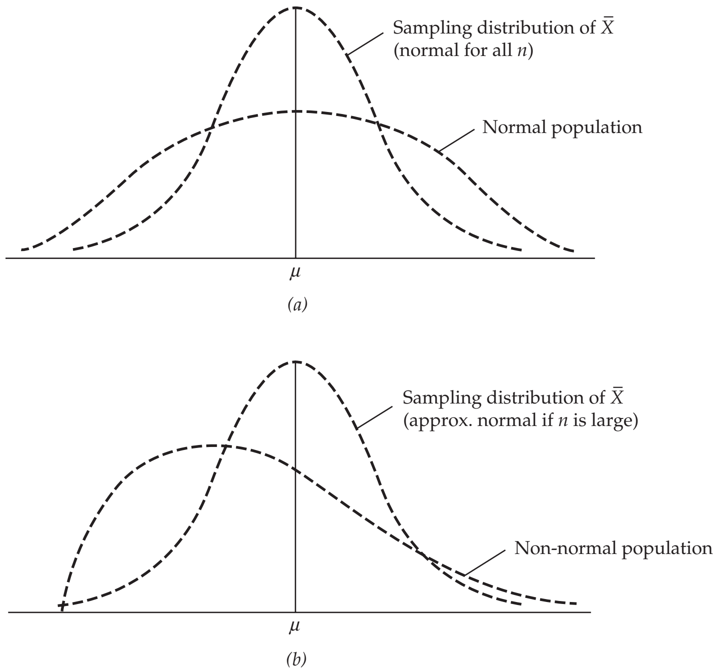

FIGURE C-5 The central limit theorem: (*a*) Samples drawn from a normal population; (*b*) samples drawn from a non-normal population

C.2 THE *t* DISTRIBUTION

The probability distribution that we use most intensively in this book is the ***t* distribution,** also known as **Student's *t* distribution.**[9] It is closely related to the normal distribution.

To introduce this distribution, recall that if $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$, the variable

$$Z = \frac{(\bar{X} - \mu_X)}{\sigma_X/\sqrt{n}} \sim N(0, 1)$$

that is, the standard normal distribution. This is so provided that both μ_X and σ_X^2 are known. But suppose we only know μ_X and estimate σ_X^2 by its (sample) estimator $S_x^2 = \frac{\Sigma(X_i - \bar{X})^2}{n - 1}$, given in Eq. (B.44). Replacing σ_X by S_x, that is, replacing the population standard deviation (s.d.) by the sample s.d., in Equation (C.7), we obtain a new variable

$$t = \frac{\bar{X} - \mu_X}{S_x/\sqrt{n}} \tag{C.8}$$

[9] *Student* was the pseudonym of W. S. Gosset, who used to work as a statistician for the Guinness Brewery in Dublin. He discovered this probability distribution in 1908.

Statistical theory shows that the t variable thus defined follows Student's t distribution with $(n-1)$ d.f. Just as the mean and variance are the parameters of the normal distribution, the t distribution has a single parameter, namely, the d.f., which in the present case are $(n-1)$. *Note:* Before we compute S_x^2 (and hence S_x), we must first compute $\overline{X}$. But since we use the same sample to compute $\overline{X}$, we have $(n-1)$, not n, independent observations to compute S^2; so to speak, we lose 1 d.f.

In sum, if we draw random samples from a normal population with mean μ_X and variance σ_X^2 but replace σ_X^2 by its estimator S_x^2, the sample mean $\overline{X}$ follows the t distribution. A t-distributed r.v. is often designated as t_k, where k denotes the d.f. (To avoid confusion with the sample size n, we use the subscript k to denote the d.f. in general.) Table E-2 in Appendix E tabulates the t distribution for various d.f. We will demonstrate the use of this table shortly.

Properties of the *t* Distribution

1. The t distribution, like the normal distribution, is symmetric, as shown in Figure C-6.
2. The mean of the t distribution, like the standard normal distribution, is zero, but its variance is $k/(k-2)$. Therefore, the variance of the t distribution is defined for d.f. greater than 2.

We have already seen that for the standard normal distribution the variance is always 1, which means that the variance of the t distribution is larger than the variance of the standard normal distribution, as shown in Figure C-6. In other words, the t distribution is flatter than the normal distribution. But as k increases, the variance of the t distribution approaches the variance of the standard normal distribution, namely, 1. Thus, if the d.f. are $k = 10$, the variance of the t distribution is $10/8 = 1.25$; if $k = 30$, the variance becomes $30/28 = 1.07$; and when $k = 100$, the variance becomes $100/98 = 1.02$, which is not much greater than 1. As a result, *the t distribution approaches the standard normal distribution as the d.f. increase.* But notice that even for k as small as 30, there is not a

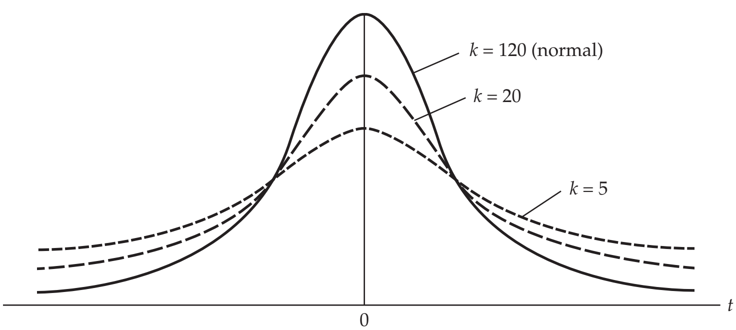

FIGURE C-6 The t distribution for selected degrees of freedom (d.f.)

great difference in the variances of the t and the standard normal variable. Therefore, the sample size does not have to be enormously large for the t distribution to approximate the normal distribution.

To illustrate the t table (Table E-2) given in Appendix E, we now consider a few examples.

Example C.8.

Let us revisit Example C.2. In a period of 15 days the sale of bread averaged 74 loaves with a (sample) s.d. of 4 loaves. What is the probability of obtaining such a sale given that the true average sale is 70 loaves a day?

If we had known the true σ, we could have used the standard normal Z variable to answer this question. But since we know its estimator, S, we can use Eq. (C.8) to compute the t value and use Table E-2 in Appendix E to answer this question as follows:

$$\begin{aligned} t &= \frac{74 - 70}{4/\sqrt{15}} \\ &= 3.873 \end{aligned}$$

Notice that in this example the d.f. are 14 = (15 – 1). (Why?)

As Table E-2 shows, for 14 d.f. the probability of obtaining a t value of 2.145 or greater is 0.025 (2.5 percent), of 2.624 or greater is 0.01 (1 percent), and of 3.787 or greater is 0.001 (0.1 percent). Therefore, the probability of obtaining a t value of as much as 3.873 or greater must be much smaller than 0.001.

Example C.9.

Let us keep the setup of Example C.8 intact except to assume that the sale of bread averages 72 loaves in the said 15-day period. Now what is the probability of obtaining such a sales figure?

Following exactly the same line of reasoning, the reader can verify that the computed t value is ~1.936. Now from Table E-2 we observe that for 14 d.f. the probability of obtaining a t value of 1.761 or greater is ~0.05 (or 5 percent) and that of 2.145 or greater is 0.025 (or 2.5 percent). Therefore, the probability of obtaining a t value of 1.936 or greater lies somewhere between 2.5 and 5 percent.

Example C.10.

Now assume that in a 15-day period the average sale of bread was 68 loaves with an s.d. of 4 loaves a day. If the true mean sales are 70 loaves a day, what is the probability of obtaining such a sales figure?

Plugging in the relevant numbers in Equation (C.8), we find that the t value in this case is –1.936. But since the t distribution is symmetric, the probability of obtaining a t value of –1.936 or smaller is the same as that of obtaining a t value of $+1.936$ or greater, which, as we saw earlier, is somewhere between 2.5 and 5 percent.

Example C.11.

Again, continue with the previous example. What is the probability that the average sale of bread in the said 15-day period was either greater than 72 loaves or less than 68 loaves?

From Examples C.9 and C.10 we know that the probability of the average sale exceeding 72 or being less than 68 is the same as the probability that a t value either exceeds 1.936 or is smaller than –1.936.[10] These probabilities, as we saw previously, are each between 0.025 and 0.05. Therefore, the total probability will be between 0.05 or 0.10 (or between 5 and 10 percent). In cases like this we would, therefore, compute the probability that $|t| > 1.936$, where $|t|$ means the absolute value of t, that is, the t value disregarding the sign. (For example, the absolute value of 2 is 2 and the absolute value of -2 is also 2.)

From the preceding examples we see that once we compute the ***t* value** from Eq. (C.8), and once we know the d.f., computing the probabilities of obtaining a given t value involves simply consulting the t table. We will consider further uses of the t table in the regression context at appropriate places in the text.

Example C.12.

For the years 1972 to 2007 the Scholastic Aptitude Test (S.A.T.) scores were as follows:

	Male	Female
Critical reading (average)	510.03	503.00
	(36.54)	(51.09)
Math (average)	524.83	486.36
	(48.31)	(102.07)

Note: The figures in parentheses are the variances.

A random sample of 10 male S.A.T. scores on the critical reading test gave the (sample) mean value of 510.12 and the (sample) variance of 41.08. What is the probability of obtaining such a score knowing that for the entire 1972–2007 period the (true) average score was 510.03?

With the knowledge of the t distribution, we can now answer this question easily. Substituting the relevant values in Eq. (C.8), we obtain

$$t = \frac{510.12 - 510.03}{\sqrt{\frac{41.08}{10}}} = 0.0444$$

[10]Be careful here. The number -2.0 is smaller than -1.936, and the number -2.3 is smaller than -2.0.

This t value has the t distribution with 9 d.f. (Why?) From Table E-2 we observe that the probability of obtaining such a t value is greater than 0.25 or 25 percent.

A note on the use of the t table (Table E-2): With the advent of user-friendly statistical software packages and electronic statistical tables, Table E-2 is now of limited value because it gives probabilities for a few selected d.f. This is also true of the other statistical tables given in Appendix E. Therefore, if you have access to one or more statistical software packages, you can compute probabilities for any given degrees of freedom much more accurately than using those given in the tables in Appendix E.

C.3 THE CHI-SQUARE (χ^2) PROBABILITY DISTRIBUTION

Now that we have derived the sampling distribution of the sample mean $\bar{X}$, (normal if the true standard deviation is known or the t distribution if we use the sample standard deviation) can we derive the sampling distribution of the sample variance, $S^2 = \frac{\sum(X_i - \bar{X})^2}{n - 1}$, since we use the sample mean and sample variance very frequently in practice? The answer is yes, and the probability distribution that we need for this purpose is the **chi-square (χ^2) probability distribution,** which is very closely related to the normal distribution. Note that just as the sample mean will vary from sample to sample, so will the sample variance. That is, like the sample mean, the sample variance is also a random variable. Of course, when we have a specific sample, we have a specific sample mean and a specific sample variance value.

We already know that if a random variable (r.v.) X follows the normal distribution with mean μ_X and variance σ_X^2, that is, $X \sim N(\mu_X, \sigma_X^2)$, then the r.v. $Z = (X - \mu_X)/\sigma_X$ is a standard normal variable, that is, $Z \sim N(0, 1)$. Statistical theory shows that the *square of a standard normal variable is distributed as a chi-square (χ^2) probability distribution with one degree of freedom (d.f.).* Symbolically,

$$Z^2 = \chi^2_{(1)} \tag{C.9}$$

where the subscript (1) of χ^2 shows the **degrees of freedom (d.f.)**—1 in the present case. As in the case of the t distribution, the d.f. is the parameter of the chi-square distribution. In Equation (C.9) there is only 1 d.f. since we are considering only the square of one standard normal variable.

A note on degrees of freedom: In general, the number of d.f. means the number of independent observations available to compute a statistic, such as the sample mean or sample variance. For example, the sample variance of an r.v. X is defined as $S^2 = \sum(X_i - \bar{X})^2/(n - 1)$. In this case we say that the number of d.f. is $(n - 1)$ because if we use the same sample to compute the sample mean $\bar{X}$, around which we measure the sample variance, so to speak, we lose one d.f.; that is, we have only $(n - 1)$ independent observations. An example will clarify this further. Consider three X values: 1, 2, and 3. The sample mean is 2. Now since $\sum(X_i - \bar{X}) = 0$ always, of the three deviations (1 – 2), (2 – 2), and (3 – 2),

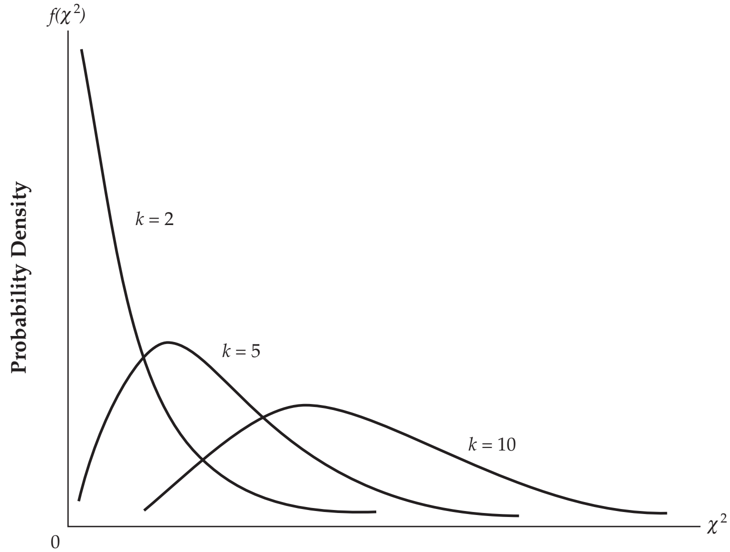

FIGURE C-7 Density function of the χ^2 variable

only two can be chosen arbitrarily; the third must be fixed in such a way that the condition $\sum(X_i - \bar{X}) = 0$ is satisfied.[11] Therefore, in this case, although there are three observations, the d.f. are only 2.

Now let $Z_1, Z_2, \ldots, Z_k$ be k *independent* unit normal variables (i.e., each Z is a normal r.v. with zero mean and unit variance). If we square each of these Z's, we can show that the sum of the squared Z's also follows a chi-square distribution with k d.f. That is,

$$\sum Z_i^2 = Z_1^2 + Z_2^2 + \cdots Z_k^2 \sim \chi^2_{(k)} \tag{C.10}$$

Note that the d.f. are now k since there are k independent observations in the sum of squares shown in Equation (C.10).

Geometrically, the χ^2 distribution appears as in Figure C-7.

Properties of the Chi-square Distribution

1. As Figure C-7 shows, unlike the normal distribution, the chi-square distribution takes only positive values (after all, it is the distribution of a squared quantity) and ranges from 0 to infinity.
2. As Figure C-7 also shows, unlike the normal distribution, the chi-square distribution is a *skewed distribution,* the degree of the skewness depending on the d.f. For comparatively few d.f. the distribution is highly skewed to

[11]Note that $\sum(X_i - \bar{X}) = \sum X_i - \sum \bar{X} = n\bar{X} - n\bar{X} = 0$, because $\bar{X} = \sum X_i/n$ and $\sum \bar{X} = n\bar{X}$, because $\bar{X}$ is a constant, given a particular sample.

the right, but as the d.f. increase, the distribution becomes increasingly symmetrical and approaches the normal distribution.

3. The expected, or mean, value of a chi-square r.v. is k and its variance is $2k$, where k is the d.f. *This is a noteworthy property of the chi-square distribution in that its variance is twice its mean value.*
4. If Z_1 and Z_2 are two *independent* chi-square variables with k_1 and k_2 d.f., then their sum $(Z_1 + Z_2)$ is also a chi-square variable with d.f. $= (k_1 + k_2)$.

Table E-4 in Appendix E tabulates the probabilities that a particular χ^2 value exceeds a given number, assuming the d.f. underlying the chi-square value are known or given. Although specific applications of the chi-square distribution in regression analysis will be considered in later chapters, for now we will look at how to use the table.

Example C.13.

For 30 d.f., what is the probability that an observed chi-square value is greater than 13.78? Or greater than 18.49? Or greater than 50.89?

From Table E-4 in Appendix E we observe that these probabilities are 0.995, 0.95, and 0.01, respectively. Thus, for 30 d.f. the probability of obtaining a chi-square value of approximately 51 is very small, only about 1 percent, but for the same d.f. the probability of obtaining a chi-square value of approximately 14 is very high, about 99.5 percent.

Example C.14.

If S^2 is the sample variance obtained from a random sample of n observations from a normal population with the variance of σ^2, statistical theory shows that the quantity

$$(n - 1)\left(\frac{S^2}{\sigma^2}\right) \sim \chi^2_{(n-1)} \tag{C.11}$$

That is, the ratio of the sample variance to population variance multiplied by the d.f. $(n - 1)$ follows the chi-square distribution with $(n - 1)$ d.f. Suppose a random sample of 20 observations from a normal population with $\sigma^2 = 8$ gave a sample variance of $S^2 = 16$. What is the probability of obtaining such a sample variance?

Putting the appropriate numbers in the preceding expression, we find that $19(16/8) = 38$ is a chi-square variable with 19 d.f. And from Table E-4 in Appendix E we find that for 19 d.f. if the true σ^2 were 8, the probability of finding a chi-square value of ≈ 38 is ≈ 0.005, a very small probability. There is doubt whether the particular random sample came from a population with a variance value of 8. But we will discuss this more in Appendix D.

In Appendix D we will show how Eq. (C.11) enables us to test hypotheses about σ^2 if we have knowledge only about the sample variance S^2.

C.4 THE *F* DISTRIBUTION

Another probability distribution that we find extremely useful in econometrics is the ***F* distribution.** The motivation behind this distribution is as follows. Let $X_1, X_2, \ldots, X_m$ be a random sample of size m from a normal population with mean μ_X and variance σ_X^2, and let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from a normal population with mean μ_y and variance σ_y^2. Assume that these two samples are independent and are drawn from populations that are normally distributed. Suppose we want to find out if the variances of the two normal populations are the same, that is, whether $\sigma_X^2 = \sigma_y^2$. Since we cannot directly observe the two population variances, let us suppose we obtain their estimators as follows:

$$S_X^2 = \sum \frac{(X_i - \overline{X})^2}{m - 1} \tag{C.12}$$

$$S_Y^2 = \sum \frac{(Y_i - \overline{Y})^2}{n - 1} \tag{C.13}$$

Now consider the following ratio:

$$\begin{aligned} F &= \frac{S_X^2}{S_Y^2} \\ &= \frac{\sum (X_i - \overline{X})^2/(m - 1)}{\sum (Y_i - \overline{Y})^2/(n - 1)} \end{aligned} \tag{C.14}$$ [12]

If the two population variances are in fact equal, the F ratio given in Equation (C.14) should be about 1, whereas if they are different, the F ratio should be different from 1; the greater the difference between the two variances, the greater the F value will be.

Statistical theory shows that if $\sigma_X^2 = \sigma_y^2$ (i.e., the two population variances are equal), the F ratio given in Eq. (C.14) follows the *F distribution* with $(m - 1)$ (numerator) d.f. and $(n - 1)$ (denominator) d.f.[13] And since the F distribution is often used to compare the variances of two (approximately normal) populations, it is also known as the **variance ratio distribution.** The F ratio is often designated

[12]By convention, in computing the F value the variance with the larger numerical value is put in the numerator. That is why the F value is always 1 or greater than 1. Also, note that if a variable, say, W, follows the F distribution with m and n d.f. in the numerator and denominator, respectively, then the variable $(1/W)$ also follows the F distribution but with n and m d.f. in the numerator and denominator, respectively. More specifically,

$$F_{(1-\alpha),m,n} = \frac{1}{F_{\alpha,n,m}}$$

where α denotes the level of significance, which we will discuss in Appendix D.

[13]To be precise, $\frac{S_x^2/\sigma_X^2}{S_y^2/\sigma_y^2}$ follows the F distribution. But if $\sigma_X^2 = \sigma_y^2$, we have the F ratio given in Eq. (C.14). Note that in computing the two sample variances we lose 1 d.f. for each, because in each case, we use the same sample to compute the sample mean, which consumes 1 d.f.

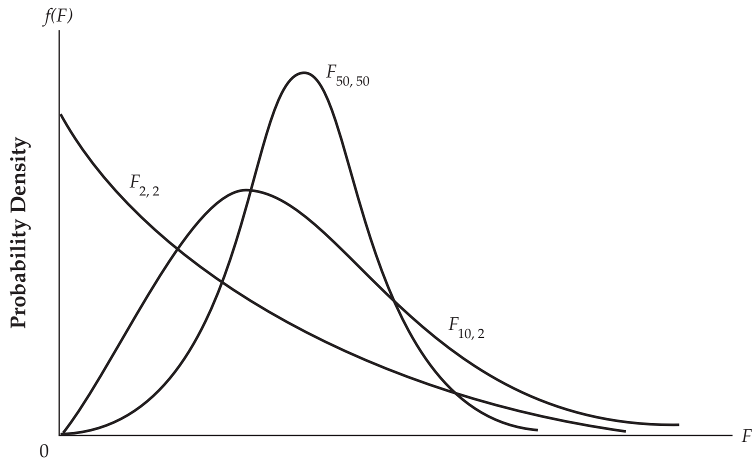

FIGURE C-8 The *F* distribution for various d.f.

as F_{k_1,k_2}, where the double subscript indicates the parameters of the distribution, namely, the **numerator and the denominator d.f.** (in the preceding example, $k_1 = [m - 1]$ and $k_2 = [n - 1]$).[14]

Properties of the *F* Distribution

1. Like the chi-square distribution, the *F* distribution is also skewed to the right and also ranges between 0 and infinity (see Figure C-8).
2. Also, like the *t* and chi-square distributions, the *F* distribution approaches the normal distribution as k_1 and k_2, the d.f., become large (technically, infinite).
3. The square of a *t*-distributed r.v. with *k* d.f. has an *F* distribution with 1 and *k* d.f. in the numerator and denominator, respectively. That is,

$$t_k^2 = F_{1,k} \tag{C.15}$$

   We will see the usefulness of this property in Chapter 4.
4. Just as there is a relationship between the *F* and *t* distributions, there is a relationship between the *F* and chi-square distributions, which is

$$F_{(m,n)} = \frac{\chi^2}{m} \qquad as\ n \rightarrow \infty \tag{C.16}$$

   That is, a chi-square variable divided by its d.f., *m*, approaches the *F* variable with *m* d.f. in the numerator and very large (technically, infinite) d.f. in the denominator. Therefore, in very large samples, we can use the

[14]The *F* distribution has two sets of d.f. because statistical theory shows that the *F* distribution is the distribution of the ratio of two *independent chi-square* random variables divided by their respective d.f.

χ^2 distribution instead of the F distribution, and vice versa. We can write Eq. (C.16) alternatively as

$$m \cdot F_{(m,n)} = \chi^2_m \qquad as\ n \to \infty \tag{C.17}$$

That is, numerator d.f. times $F_{(m,n)}$ equals a chi-square value with numerator d.f., provided the denominator degrees of freedom are sufficiently large (technically, infinite).

The F distribution is tabulated in Table E-3 in Appendix E. We will consider its specific uses in the context of regression analysis in the text, but in the meantime let us see how this table is used.

Example C.15.

Let us return to the S.A.T. example (Example C.12). Assume that the critical reading scores for males and females are each normally distributed. Further assume that average scores and their variances given in the preceding table represent sample values from a much larger population. Based on the two sample variances, can we assume that the two population variances are the same?

Since the critical reading scores of the male and female populations are assumed to be normally distributed random variables, we can compute the F ratio given in Eq. (C.14) as

$$F = \frac{51.09}{36.54} = 1.3982$$

which has the F distribution with 35 d.f. in the numerator and 35 d.f. in the denominator. (*Note:* In computing the F value, we are putting the larger of the two variances in the numerator.) Although Table E-3 in Appendix E does not give the F value corresponding to d.f. of 35, if we use 30 d.f. for both the numerator and the denominator, the probability of obtaining an F value of about 1.40 lies somewhere between 10 and 25 percent. Since this probability is not very low (more about this in Appendix D), we could say there does not seem to be enough evidence to claim the two population variances are unequal. Therefore, we decide there is not a difference in the population variances of male and female scores on the critical reading part of the S.A.T. test. Remember that if the two population variances are the same, the F value will be 1, but if they are different, the F value will be increasingly greater than 1.

Example C.16.

An instructor gives the same econometrics examination to two classes, one consisting of 100 students and the other consisting of 150 students. He draws a random sample of 25 students from the first class and a random sample of 31 students from the other class and observes that the sample variances of the grade-point average in the two classes are 100 and 132, respectively. It is assumed that the r.v., grade-point average, in the two classes is normally distributed. Can we assume that the variances of grade-point average in the two classes are the same?

Since we are dealing with two independent random samples drawn from two normal populations, applying the F ratio given in Eq. (C.14), we find that

$$F = \frac{132}{100} = 1.32$$

follows the F distribution with 30 and 24 d.f., respectively. From the F values given in Table E-3 we observe that for 30 numerator d.f. and 24 denominator d.f. the probability of obtaining an F value of as much as 1.31 or greater is 25 percent. If we regard this probability as reasonably high, we can conclude that the (population) variances in the two econometrics classes are (statistically) the same.

C.5 SUMMARY

In Appendix A we discussed probability distributions in general terms. In this appendix, we considered four specific probability distributions—the normal, the t, the chi-square, and the F—and the special features of each distribution, in particular, the situations in which these distributions can be useful. As we will see in the main chapters of this book, these four PDFs play a very pivotal role in econometric theory and practice. Therefore, a solid grasp of the fundamentals of these distributions is essential to follow the text material. You may want to return to this appendix from time to time to consult specific points of these distributions when they are referred to in the main chapters.

KEY TERMS AND CONCEPTS

The key terms and concepts introduced in this appendix are

The normal distribution
- **a)** unit or standardized variable
- **b)** unit or standard normal variable

Random number generators
Random sampling; i.i.d. random variables
Sampling, or probability, distribution of an estimator (e.g., the sample mean)
Standard error (se)
Monte Carlo experiments or simulations
Central limit theorem (CLT)
t distribution (Student's t distribution)
Chi-square (χ^2) probability distribution
Degrees of freedom (d.f.)
F distribution
- **a)** variance ratio distribution
- **b)** numerator and denominator degrees of freedom (d.f.)

QUESTIONS

C.1. Explain the meaning of
  a. Degrees of freedom.
  b. Sampling distribution of an estimator.
  c. Standard error.

C.2. Consider a random variable (r.v.) $X \sim N(8, 16)$. State whether the following statements are true or false:
  a. The probability of obtaining an X value of greater than 12 is about 0.16.
  b. The probability of obtaining an X value between 12 and 14 is about 0.09.
  c. The probability that an X value is more than 2.5 standard deviations from the mean value is 0.0062.

C.3. Continue with Question C.2.
  a. What is the probability distribution of the sample mean $\overline{X}$ obtained from a random sample from this population?
  b. Does your answer to (*a*) depend on the sample size? Why or why not?
  c. Assuming a sample size of 25, what is the probability of obtaining an $\overline{X}$ of 6?

C.4. What is the difference between the t distribution and the normal distribution? When should you use the t distribution?

C.5. Consider an r.v. that follows the t distribution.
  a. For 20 degrees of freedom (d.f.), what is the probability that the t value will be greater than 1.325?
  b. What is the probability that the t value in C.5(*a*) will be less than -1.325?
  c. What is the probability that a t value will be greater than or less than 1.325?
  d. Is there a difference between the statement in C.5(*c*) and the statement, "What is the probability that the absolute value of t, $|t|$, will be greater than 1.325?"

C.6. *True or false*. For a sufficiently large d.f., the t, the chi-square, and the F distributions all approach the unit normal distribution.

C.7. For a sufficiently large d.f., the chi-square distribution can be approximated by the standard normal distribution as: $Z = \sqrt{2\chi^2} - \sqrt{2k - 1} \sim N(0, 1)$. Let $k = 50$.
  a. Use the chi-square table to find out the probability that a chi-square value will exceed 80.
  b. Determine this probability by using the preceding normal approximation.
  c. Assume that the d.f. are now 100. Compute the probability from the chi-square table as well as from the given normal approximation. What conclusions can you draw from using the normal approximation to the chi-square distribution?

C.8. What is the importance of the central limit theorem in statistics?

C.9. Give examples where the chi-square and F probability distributions can be used.

PROBLEMS

C.10. Profits (X) in an industry consisting of 100 firms are normally distributed with a mean value of \$1.5 million and a standard deviation (s.d.) of \$120,000. Calculate
  a. $P(X < \$1 \text{ million})$
  b. $P(\$800{,}000 \le X \le \$1{,}300{,}000)$

C.11. In Problem C.10, if 10 percent of the firms are to exceed a certain profit, what is that profit?

C.12. The grade-point average in an econometrics examination was normally distributed with a mean of 75. In a sample of 10 percent of students it was found that the grade-point average was greater than 80. Can you tell what the s.d. of the grade-point average was?

C.13. The amount of toothpaste in a tube is normally distributed with a mean of 6.5 ounces and an s.d. of 0.8 ounces. The cost of producing each tube is 50 cents. If in a quality control examination a tube is found to weigh less than 6 ounces, it is to be refilled to the mean value at a cost of 20 cents per tube. On the other hand, if the tube weighs more than 7 ounces, the company loses a profit of 5 cents per tube.

If 1000 tubes are examined,

a. How many tubes will be found to contain less than 6 ounces?

b. In that case, what will be the total cost of the refill?

c. How many tubes will be found to contain more than 7 ounces? In that case, what will be the amount of profits lost?

C.14. If $X \sim N(10, 3)$ and $Y \sim N(15, 8)$, and if X and Y are independent, what is the probability distribution of

a. $X + Y$ **b.** $X - Y$ **c.** $3X$ **d.** $4X + 5Y$

C.15. Continue with Problem C.14, but now assume that X and Y are positively correlated with a correlation coefficient of 0.6.

C.16. Let X and Y represent the rates of return (in percent) on two stocks. You are told that $X \sim N(15, 25)$ and $Y \sim N(8, 4)$, and that the correlation coefficient between the two rates of return is -0.4. Suppose you want to hold the two stocks in your portfolio in equal proportion. What is the probability distribution of the return on the portfolio? Is it better to hold this portfolio or to invest in only one of the two stocks? Why?

C.17. Return to Example C.12. A random sample of 10 female S.A.T. scores on the math test gave a sample variance of 142. Knowing that the true variance is 102.07, what is the probability of obtaining such a sample value? Which probability distribution will you use to answer this question? What are the assumptions underlying that distribution?

C.18. The 10 economic forecasters of a random sample were asked to forecast the rate of growth of the real gross national product (GNP) for the coming year. Suppose the probability distribution of the r.v.—forecast—is normal.

a. The probability is 0.10 that the sample variance of the forecast is more than X percent of the population variance. What is the value of X?

b. If the probability is 0.95 so that the sample variance is between X and Y percent of the population variance, what will be the values of X and Y?

C.19. When a sample of 10 cereal boxes of a well-known brand was reweighed, it gave the following weights (in ounces):

16.13	16.02	15.90	15.83	16.00
15.79	16.01	16.04	15.96	16.20

a. What is the sample mean? And the sample variance?

b. If the true mean weight per box was 16 ounces, what is the probability of obtaining such a (sample) mean? Which probability distribution did you use and why?

C.20. The same microeconomics examination was given to students at two different universities. The results were as follows:

$$\bar{X}_1 = 75, \qquad S_1^2 = 9.0, \qquad n_1 = 50$$
$$\bar{X}_2 = 70, \qquad S_2^2 = 7.2, \qquad n_2 = 40$$

where the $\bar{X}$'s denote the grade averages in the two samples, the S^2's, the two sample variances; and the n's, the sample sizes. How would you test the hypothesis that the population variances of the test scores in the two universities are the same? Which probability distribution would you use? What are the assumptions underlying that distribution?

C.21. *Monte Carlo Simulation.* Draw 25 random samples of 25 observations from the t distribution with $k = 10$ d.f. For each sample compute the sample mean. What is the sampling distribution of these sample means? Why? You may use graphs to illustrate your answer.

C.22. Repeat Problem C.21, but this time use the χ^2 distribution with 8 d.f.

C.23. Repeat Problem C.21, but use the F distribution with 10 and 15 d.f. in the numerator and denominator, respectively.

C.24. Using Eq. (C.16), compare the values of $\chi^2_{(10)}$ with $F_{10,10}$, $F_{10,20}$, and $F_{10,60}$. What general conclusions do you draw?

C.25. Given $X \sim N(\mu_X, \sigma_X^2)$, prove that $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$. *Hint:* $\text{var}(\bar{X}) = \text{var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$. Expand this, recalling some of the properties of the variance discussed in Appendix B and the fact that the X_i are i.i.d.

C.26. Prove that $Z = \left(\frac{X - \mu_X}{\sigma_X}\right)$, has zero mean and unit variance. Note that this is true whether Z is normal or not. *Hint:* $E(Z) = E\left(\frac{X - \mu_X}{\sigma_X}\right) = \frac{1}{\sigma_X} E(X - \mu_X)$.