CHAPTER 4

Sampling Distributions

This chapter introduces the reader the basic sampling distributions. Inferential statistics is used here to draw conclusions about a population, based on a probabilistic model of random samples of the population. Since different random samples will most likely give different estimates, some knowledge of the variability of all possible estimates derived from random samples is needed to arrive at reasonable conclusions. *Population* is any finite set of objects being investigated. A sample of objects drawn from a population is a *random sample*.

If the populations contains N elements and a sample of n of them is to be selected, then if each of the $\frac{N!}{(N-n)!n!}$ possible samples has an equal probability of being chosen, the procedure employed is known as *random sampling*. Due to the difficulty in obtaining random samples in practice tables of random numbers are used.

Sampling theory is the study of relationships existing between a population and samples drawn from the population. Sampling theory is useful in finding whether observed differences between two samples are actually due to chance variation or whether they are really significant. A study of inferences made regarding a population by the use of samples drawn from it, along with indications of the accuracy of such inferences using probability theory is known as *statistical inferences*. A *statistic* is a function of the observations in a random sample, which is not dependent on unknown parameters. A *parameter* is in general an unknown constant. For example, the parameters of a normal distribution are μ and σ^2, whereas $\bar{X}$ and s^2 are statistics. The behaviour of sample statistics is needed in order to draw conclusions about a sample. The probabilistic distribution of a random variable defined on a space of random samples is called a *sampling distribution*. The behaviour of the sample statistics is described by a *sampling distribution*. Several sampling distributions are discussed in this chapter and their application to inferential statistics in Chapter 5 (Estimation) and Chapter 6 (Hypothesis Testing).

Suppose that $y_1, y_2, \ldots, y_n$ represents a sample. Then the *sample mean*

$$\bar{y} = \frac{\sum_{t=1}^{n} y_i}{n} \tag{4.1}$$

and the *sample variance*

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{4.2}$$

are both statistics. These quantities are measures of the central tendency and dispersion of the sample, respectively. Sometimes $s = \sqrt{s^2}$, called the *sample standard deviation*, is used as a measure of dispersion.

4.1 PROPERTIES OF SAMPLE MEAN AND VARIANCE

The sample mean $\bar{y}$ is a point estimator of the population mean μ, and the sample variance s^2 is a *point estimator* of the population variance σ^2. An *estimator* of an unknown parameter is a statistic that corresponds to that parameter. Note that a point estimator is a random variable. A numerical value of an estimator computed from sample data, is called an *estimate*. There are several properties required of good *point estimators*. Two of the most important are the following:

1. The point estimator must be *unbiased*. The long-run average of expected value of the point estimator should be the parameter that is being estimated. An unbiasedness is a desirable property and this alone does not always make an estimator a good one.
2. An unbiased estimator must have *minimum variance*. The minimum variance point estimator has a variance that is smaller than the variance of any other estimator of that parameter. Here, we show that $\bar{y}$ and s^2 are unbiased estimators of μ and σ^2, respectively. Using the properties of expectation,

$$E(\bar{y}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \tag{4.3}$$

since the expected value of each observation y_i is μ. Thus, $\bar{y}$ is an unbiased estimator of μ.

Considering the sample variances s^2. We get

$$E(S^2) = E\left[\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] = \frac{1}{n-1}E(SS) \tag{4.4}$$

where $SS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *correct sum of squares* of the observation y_i. Now

$$E(SS) = E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] = E\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right]$$

$$\sum_{i=1}^{n}(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) = (n-1)\sigma^2 \tag{4.5}$$

Hence, $$E(S^2) = \frac{1}{n-1} E(SS) = \sigma^2 \tag{4.6}$$

Therefore, s^2 is an unbiased estimator of σ^2. The quantity $n - 1$ in Equation (4.5) is called the *number of degrees of freedom* of the sum of squares *SS*. If y is a random variable with variance σ^2 and $SS = \sum (y_i - \bar{y})^2$ has degrees of freedom, then

$$E\left(\frac{SS}{\nu}\right) = \sigma^2 \tag{4.7}$$

The number of degrees of freedom of a sum of squares is equal to the number of independent elements in that sum of squares. For example, $SS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ in Equation (4.5) consists of the sum of squares of the n elements $y_1 - \bar{y}$, $y_2 - \bar{y}$, …, $y_n - \bar{y}$. These elements are not all independent since $SS = \sum_{i=1}^{n} (y_i - \bar{y}) = 0$; in fact, only $n - 1$ of them are independent, implying that *SS* has $n - 1$ degrees of freedom.

The *standard error* of a statistic is the standard deviation of its sampling distribution. If the standard error involves unknown parameters whose values can be estimated, substitution of these estimates into the standard error results in an *estimated standard error*. Suppose, we are sampling from a normal distribution with mean μ and variance σ^2. Now the distribution of $\bar{X}$ is normal with mean μ and variance σ^2/n, so the *standard error* of $\bar{X}$ is

$$\frac{\sigma}{\sqrt{n}} \tag{4.7a}$$

The *estimated standard error* of $\bar{X}$ is

$$\frac{s}{\sqrt{n}} \tag{4.7b}$$

4.2 POPULATION AND SAMPLING DISTRIBUTIONS

Here, we introduce the basic concept of population distribution and sampling distribution.

4.2.1 Population Distribution

The *population distribution* is the probability distribution of the population data. It is the probability distribution derived from the information on all elements of a population.

4.2.2 Sampling Distribution

A sample provides data for only a portion of an entire population and therefore we cannot expect the sample to yield perfectly accurate information about the population. The value of a population parameter is always constant. Any population data set has only one value of the population mean, μ. We would expect different samples of the same size drawn from the same population to give different values of the sample mean, $\bar{x}$.

The sample mean, $\bar{x}$, is therefore a *random variable* and it possesses a probability distribution. For each sample we can find a statistic, such as the mean, standard deviation, etc., which will vary from sample to sample. As a consequence, we obtain a distribution of the statistic which is called its *sampling distribution*. Sample statistics such as the mean, median, mode and standard deviation all possess sampling distributions. In general, the probability distribution of a sample statistic is called its sampling distribution. The probability distribution of $\bar{x}$ is called its sampling distribution. It lists the values that $\bar{x}$ can assume and the probability of each value of $\bar{x}$.

4.3 SAMPLING AND NONSAMPLING ERRORS

Since a sample provides data for only a portion of an entire population, we cannot expect the sample to give perfectly accurate information about the population. Hence, we can expect a certain information about the population. Hence, we can expect a certain amount of error called *sampling error* - will result simply because we are sampling. *Sampling error* is the difference between the value of a sample statistic and the value of the corresponding population parameter. In the case of the mean,

$$\text{Sampling error} = \bar{x} - \mu \tag{4.8}$$

assuming that the sample is random and no nonsampling error has been made.

A sampling error occurs due to chance. The errors that occur in the collection, recording and tabulation of data are called *nonsampling errors*. Such errors occur due to human mistakes and not chance. The larger the sample size, the smaller the sampling error tends to be in estimating a population mean, μ, by a sample mean $\bar{x}$.

Example E4.1

The mean ages of all students in a large university follow a distribution that is skewed to the right is 24 years and a standard deviation of 4 years. Find the probability that the mean age for a random sample of 36 students would be

(a) between 23 and 25 years

(b) less than 23 years.

SOLUTION:

Given the population mean = 24 years, and $n = 36$.

The standard deviation of the sample mean is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{36}} = 0.6667 \text{ years}$$

(a) For $\bar{x} = 23$: $z = \dfrac{23 - 24}{0.6667} = -1.50$

For $\bar{x} = 25$: $z = \dfrac{25 - 24}{0.6667} = 1.50$

$$P(23 < \bar{x} < 25) = P(-1.50 < z < 1.50) = (0.933193 - 0.066807) = 0.86639$$

(from the table in Appendix-*E*)

(b) For $\bar{x} = 23$: $z = \dfrac{23 - 24}{0.6667} = -1.50$

$P(\bar{x} < 25) = P(z < -1.50) = 0.066807$ (from the table in Appendix-*E*).

4.4 MEAN AND STANDARD DEVIATION OF x

For a variable x and a given sample size, the distribution of the variable $\bar{x}$ is called *the sampling distribution of the sample mean*.

The mean and standard deviation of the sampling distribution of $\bar{x}$ are called the *mean and standard deviation of* $\bar{x}$ and are denoted by $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ respectively.

The standard deviation of $\bar{x}$ is also called the *standard error of* $\bar{x}$. There is a simple relationship between the mean of the variable $\bar{x}$ and the mean of the variable under consideration. They are equal, or $\mu_{\bar{x}} = \mu$. Thus, for any particular sample size, the mean of all possible samples means equals the populations mean. This equality holds true regardless of the size of the sample.

Hence, the mean of the sampling distribution of $\bar{x}$ is always equal to the mean of the population.

$$\mu_{\bar{x}} = \mu \tag{4.9}$$

The sample mean $\bar{x}$, is called an *estimate* of the population mean, μ. When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, that simple statistic is said to be an *unbiased estimator*. For the sample mean $\bar{x}$, $\mu_{\bar{x}} = \mu$. Therefore, $\bar{x}$ is an unbiased estimator of μ.

Standard Deviation of the Sample Mean

For samples of size n, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration divided by the square root of the sample size.

Hence, $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{4.10}$$

The above Eq. (4.10) assumes the sampling is done with replacement from a finite population or when it is done from an infinite population. When sampling is done without replacement from a finite population, the appropriate formula is

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}} \tag{4.11}$$

where n denotes the sample size and N the population size.

The sample size is considered to be small compared to the population size if the sample is equal to or less than 5% of the population size.

That is, if $n/N \leq 0.05$.

When the sample size is small relative to the population size, there is little difference between sampling with and without replacement. In such cases, the two formulae Eqs. (4.10) and (4.11) for $\sigma_{\bar{x}}$ yield almost

the same values. However, in most practical applications, the sample size is small relative to the population size.

The possible sample means cluster more closely around the population mean as the sample size increases, and therefore the larger the sample size, the smaller the sampling error tends to be in estimating a population mean by a sample mean.

The spread of the sampling distribution of $\bar{x}$ is smaller than the spread of the corresponding population distribution. In other words, $\sigma_{\bar{x}} < \sigma$. The standard deviation of the sampling distribution of $\bar{x}$ decreases as the sample size increases.

If the standard deviation of a sample statistic decreases as the sample size is increased, that statistic is said to be a *consistent estimator*. From Eq.(4.10), $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, it is clear that as n increases, the value of $\sqrt{n}$ also increases and as a result, the value of $\frac{\sigma}{\sqrt{n}}$ decreases. Hence, the sample mean $\bar{x}$ is a consistent estimator of the population mean, μ.

Example E4.2

A population random variable X has mean 120 and standard deviation 15. Find the mean and standard deviation of the sample mean $\bar{X}$ for random samples of size 5 drawn with replacement.

SOLUTION:

For the population, $\mu = 120$, $\sigma = 15$.

The mean $\mu_{\bar{X}}$ and the standard deviation $\sigma_{\bar{X}}$ of $\bar{X}$ are given by

$$\mu_{\bar{X}} = \mu = 120$$

and $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{5}} = 6.7082$$

Example E4.3

A population random variable X has mean 120 and standard deviation 15. Find the mean and standard deviation of $\bar{X}$ if the population size is 300 and the samples of size 5 are drawn without replacement.

SOLUTION:

Here $N = 300$ and $n = 5$

Mean $\mu_{\bar{X}} = \mu = 120, \sigma = 15.$

and the standard deviation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{15}{\sqrt{5}}\sqrt{\frac{300-5}{300-1}} = 6.6632$$

Example E4.4

Let $S = \{1, 3, 7, 9\}$. Find the probability distribution of the sample mean $\bar{x}$ for random samples of size 2 drawn with replacement.

SOLUTION:

Since S has 4 elements, there are $4^2 = 16$ random samples of size 2 drawn with replacement. These pairs and their average values are given in Table P4.4.

Table P4.4

Sample	$\bar{x}$	Sample	$\bar{x}$	Sample	$\bar{x}$	Sample	$\bar{x}$
(1, 1)	1	(1, 3)	2	(1, 7)	4	(1, 9)	5
(3, 1)	2	(3, 3)	3	(3, 7)	5	(3, 9)	6
(7, 1)	4	(7, 3)	5	(7, 7)	7	(7, 9)	8
(9, 1)	5	(9, 3)	6	(9, 7)	8	(9, 9)	9

The probability distribution of $\bar{x}$ is given in Table P4.4(a)

Table P4.4(a)

$\bar{x}$	1	2	3	4	5	6	7	8	9
$p(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{4}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Example E4.5

Let $S = \{1, 3, 7, 9\}$

(a) list all samples of size 3, without replacement

(b) how many samples, without replacement, are there of size 4, size n?

SOLUTION:

(a) A sample of size 3, without replacement is a subset of S containing 3 elements. There are

$$\binom{4}{3} = 4$$

Subsets: $\{1, 3, 7\}$, $\{1, 3, 9\}$, $\{1, 7, 9\}$, $\{3, 7, 9\}$.

(b) For $n = 1, 2, 3, 4$, there are $\binom{4}{n}$ samples of size n; for $n > 4$, there are no samples of size n.

Example E4.6

Let $S = \{1, 3, 7, 9\}$. Find the probability distribution of the sample mean $\bar{x}$ for random samples of size 2 drawn without replacement.

SOLUTION:

Since S has 4 elements, there are $\binom{4}{2} = 6$ random samples of size 2 drawn without replacement. These, their average value, and the probability distribution of $\bar{x}$ are given in Tables P4.6 and P4.6(a).

Table P4.6

Sample	$\bar{x}$
(1, 3)	2
(1, 7)	4
(1, 9)	5
(3, 7)	5
(3, 9)	6
(7, 9)	8

Table P4.6(a)

$\bar{x}$	$p(\bar{x})$
2	1/5
4	1/6
5	2/6
6	1/6
8	1/6

Example E4.7

How many teams of 5 students can be randomly selected from a class of 10 female students and 15 male students? *(a)* how many teams will have all male students? *(b)* how many teams will have all female students? *(c)* how many teams will have 3 female students and 2 male students?

SOLUTION:

The number of 5-student teams is the number of ways at 5 students can be selected from a class of 25 students, or the number of samples of size 5 that can be selected, without replacement, from a population of size 25, which is

$$\binom{25}{5} = 53,130 \text{ (from the table in Appendix-}B\text{)}$$

(a) The number of teams that will have all male students is

$$\binom{15}{5} = 3003 \text{ (from the table in Appendix-}B\text{)}$$

(b) The number of teams that have all female students is

$$\binom{10}{5} = 252 \text{ (from the table in Appendix-}B\text{)}$$

(c) The number of teams that have 3 female students and 2 male students is

$$\binom{10}{3}\binom{15}{2} = (120)(150) = 12,600 \text{ (see Appendix-}B\text{).}$$

Example E4.8

Determine the most likely breakdown of male students and female students in a team of 5 randomly selected from 15 male and 10 female students.

SOLUTION:

The ratio of 15 male students to 10 female students is 3 to 2. Hence, a team of 3 male and 2 female students would occur at random. From Problem P4.7, we have

5 male students in a team = 3003

5 female students in a team = 252

3 female students and 2 male students in a team = 12600

In a similar manner, we obtain the following counts: (use the table in Appendix-*B*)

$$\text{1 male and 4 female students in a team} = \binom{15}{1}\binom{10}{4} = (15)(210) = 3150$$

$$\text{3 male and 2 female students in a team} = \binom{15}{3}\binom{10}{2} = (455)(45) = 20{,}475$$

$$\text{4 male and 1 female students in a team} = \binom{15}{4}\binom{10}{1} = (1365)(10) = 13{,}650$$

4.5 SHAPE OF THE SAMPLING DISTRIBUTION OF $\bar{x}$

In Section 4.4, we described the sampling distribution of the sample mean, that is, the distribution of the variable $\bar{x}$. It was shown there that the mean and standard deviation of $\bar{x}$ can be expressed in terms of the sample size and the population mean and standard deviation:

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

In this section, we describe the shape of the sampling distribution of $\bar{x}$ as related to the following two cases:

1. The population from which samples are drawn has a normal distribution.
2. The population from which samples are drawn does not have a normal distribution.

4.5.1 Sampling from a Normally Distributed Population

If the population variable x of a population is normally distributed with mean μ and standard deviation σ, then the sampling distribution of the sample mean, $\bar{x}$, will also be normally distributed with the mean μ and standard deviation $\sigma/\sqrt{n}$.

That is $\mu_{\bar{x}} = \mu$

and $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \frac{n}{N} \le 0.05 \tag{4.12}$$

From the sampling distribution of $\bar{x}$ for different sample sizes, n, it can be observed that the spread of the sampling distribution of $\bar{x}$ decreases as the sample size increases.

Example E4.9

The lengths of all machine parts made by a company have a distribution that is skewed to the right with a mean of 68 mm and a standard deviation of 4 mm. Find the probability that the mean length of a random sample of 100 parts produced by this company would be

(a) less than 67.8 mm

(b) between 67.5 mm and 68.7 mm

(c) within 0.6 mm of the population mean

(d) lower than the population mean by 0.5 mm or more.

SOLUTION:

Given $\mu = 68$ mm, $\sigma = 4$ mm and $n = 100$.

The standard deviation of the sample mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{100}} = 0.4 \text{ mm}$$

(a) For $\bar{x} = 67.8$ mm: $z = \dfrac{67.8 - 68}{0.4} = -0.50$

$P(\bar{x} < 67.8) = P(z < -0.50) = 0.308538$ (from the table in Appendix-*E*).

(b) For $\bar{x} = 67.5$: $z = \dfrac{67.5 - 68}{0.4} = -1.25$

For $\bar{x} = 68.7$: $z = \dfrac{68.7 - 68}{0.4} = 1.75$

$P(67.5 \le \bar{x} \le 67.8) = P(-1.25 \le z \le 1.75) = (-0.105650 + 0.959941) = 0.85429$

(from the table in Appendix-*E*).

(c) $P(\bar{x}$ within 0.6 mm of $\mu) = P(67.4 \le \bar{x} \le 68.6)$

For $\bar{x} = 67.4$: $z = \dfrac{67.4 - 68}{0.4} = -1.50$

For $\bar{x} = 68.6$: $z = \dfrac{68.6 - 68}{0.4} = 1.50$

$P(67.4 \le \bar{x} \le 68.6) = P(-1.50 \le z \le 1.50) = 0.933193 - 0.066807 = 0.86639$

(from the table in Appendix-*E*).

(d) $P(\bar{x}$ lower than μ by 0.5 mm or more$) = p(\bar{x} \le 67.5)$

$P(\bar{x} \le 67.5) = P(z \le -1.25) = 0.10565$ (from the table in Appendix-*E*).

For $\bar{x} = 67.5$: $z = \dfrac{67.5 - 68}{0.4} = -1.25$

4.5.2 Sampling from a Population that is not Normally Distributed

If the sampling is done from a population that is not normally distributed, then the shape of the sampling distribution of $\bar{x}$ is inferred from central limit theorem (see Chapter 3, Section 3.8).

4.6 APPLICATIONS OF THE SAMPLING DISTRIBUTION OF $\bar{x}$

It was shown in Chapter 3 (Section 3.8) on central limit theorem that for large samples, the sampling distribution of $\bar{x}$ is approximately normal, regardless of the distribution of the variable under considerations. The approximation becomes better with increasing sample size. In general, the farther the variable under consideration is from being normally distributed, the larger the sample size must be for a normal distribution to provide an adequate approximation to the distribution of $\bar{x}$. A sample size of 30 or more ($n \geq 30$) is large enough.

The sampling distribution of the sample mean can be summarised as follows:

If the variable x of a population has mean μ and standard deviation σ, then for samples of size n:

1. The mean of $\bar{x}$ equals the population mean, or $\mu_{\bar{x}} = \mu$.
2. The standard deviation of $\bar{x}$ equals the population standard deviation divided by the square root of the sample size, or $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.
3. If x is normally distributed, so is $\bar{x}$, regardless of sample size.
4. If the sample size is large, $\bar{x}$ is approximately normally distributed, regardless of the distribution of x.

Example E4.10

Refer to Problem P4.2. Suppose the random variable X in Problem P4.2 is approximately normally distributed, determine $P(115 \leq \bar{X} \leq 125)$ for samples of size 5 drawn with replacement.

SOLUTION:

From the solution of P4.2, the mean and standard deviation of $\bar{X}$ are $\mu_{\bar{x}} = 120$ and $\sigma_{\bar{x}} = 6.7082$. $\bar{X}$ is approximately normally distributed. Hence,

$$P(115 \leq \bar{X} \leq 125) = P\left(\frac{115-120}{6.7082} \leq \frac{\bar{X}-120}{6.7082} \leq \frac{125-120}{6.7082}\right)$$

$$= P(-0.745 \leq z \leq 0.745) \qquad \text{(See the table in Appendix-}E\text{).}$$

Where z is the standard normal random variable.

$$= (-0.229650 + 0.770350) = 0.5407$$

Example E4.11

Refer to Problem P4.3. Suppose the random variable X in Problem P4.3 is approximately normally distributed, find $P(115 \leq \bar{X} \leq 125)$ for samples of size 5 drawn with replacement.

SOLUTION:

Refer to the solution of Problem P4.3. The mean and standard deviation of $\bar{X}$ are $\mu_{\bar{x}} = 120$ and $\sigma_{\bar{x}} = 6.66.2$.

$\bar{X}$ is approximately normally distributed. Hence,

$$P(115 \le \bar{X} \le 125) = P\left(\frac{115-120}{6.6632} \le \frac{\bar{X}-120}{6.6632} \le \frac{125-120}{6.6632}\right) = P(-0.75039 \le z \le 0.75039)$$

where z is the standard normal random variable.

Using the standard normal table in Appendix-*E*, we have

$$P(-0.75039 \le z \le 0.75039) = (-0.226627 + 0.773373) = 0.54675$$

4.7 POPULATION AND SAMPLE PROPORTIONS

The *population proportion*, denoted by p, is obtained by taking the ratio of the number of elements in a population with a specific characteristic to the total number of elements in the population. In general, when we pick a sample,

$$\text{Sample proportion of an attribute} = \frac{\text{number of items in the sample having the attribute}}{\text{sample size}} \tag{4.13}$$

Hence, if X is the number of items having a certain attribute in a sample of size n, then *the sample proportion having the attribute* is the random variable X/n. The probability distribution of this statistic is called the *sampling distribution of the proportion*.

The *population* and *sample proportions*, denoted by p and $\hat{p}$, respectively, are calculated as

$$p = \frac{X}{N} \quad \text{and} \quad \hat{p} = \frac{x}{n} \tag{4.14}$$

where N = total number of elements in the population

n = total number of elements in the sample

X = number of elements in the population that possess a specific characteristic

x = number of elements in the sample that possess a specific characteristic

The sampled population can be described by the following probability distribution:

x	Probability
0	1 – p
1	p

The computations for finding the mean and the variance of this population are shown in Table 4.1.

Table 4.1

Value x	Probability p(x)	xp(x)	$x^2p(x)$
0	1 – p	0(1 – p)	$0^2(1-p)$
1	p	1p	1^2p
Σ	1	P	p

From column 3 in Table 4.1, $\Sigma xp(x) = p$.

Hence, the population mean μ is p or $\mu = p$ (4.15)

Similarly the sum in column 4 is $\Sigma x^2 p(x) = p$.

Hence, the variance of the population, σ^2 is given by

$\sigma^2 = \Sigma x^2 p(x) - \mu^2 = p - p^2 = p(1 - p)$

Therefore, population standard deviation = $\sqrt{p(1-p)}$ (4.16)

4.8 SAMPLING DISTRIBUTION OF $\hat{p}$

Similar to the sample mean $\bar{x}$, the sample proportion, $\hat{p}$, is also a random variable. It possesses a probability distribution called its *sampling distribution*. It gives the various values that $\hat{p}$ can assume and their probabilities.

4.9 MEAN AND STANDARD DEVIATION OF $\hat{p}$

The *mean of* $\hat{p}$, which is the same as the mean of the sampling distribution $\hat{p}$, is always equal to the population proportion, p.

Hence, $\mu_{\hat{p}} = p$ (4.17)

The sample proportion, $\hat{p}$, is called an *estimator* of the population proportion, p. When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, that sample statistic is said to be an *unbiased estimator*. Since for the sample population, $\mu_{\hat{p}} = p$, $\hat{p}$ is an unbiased estimator of p. The *standard deviation of* $\hat{p}$, denoted by $\sigma_{\hat{p}}$, is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} \tag{4.18}$$

where p is the population proportion, $q = 1 - p$, and n is the sample size.

For large values of n ($n \geq 30$) the sampling distribution is very closely normally distributed. Note that the population is binomially distributed. The Eq. (4.18) is also valid for a finite population in which sampling is with replacement.

However, if $n/N \geq 0.05$, then $\sigma_{\hat{p}}$ is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{N-n}{N-1}} \tag{4.19}$$

where the factor $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction factor.

In general, the sample size, n is small compared to the population size, N and Eq.(4.18) is used.

If the standard deviation of a sample statistic decreases as the sample size is increased, that statistic is called the *consistent estimator*. It is clear from Eq.(4.18) that as n increases, the value of $\sqrt{pq/n}$ decreases and the sample proportion, $\hat{p}$, is said to be the *consistent estimator* of the population proportion, p.

The shape of the sampling distribution of $\hat{p}$ is inferred from the central limit theorem described in Section 3.8 of Chapter 3. According to the central limit theorem, the *sampling distribution of* $\hat{p}$ is approximately normal for a sufficiently large sample size. In the case of a proportions, the sample size is considered to be sufficiently large if np and $n(1-p)$ are both greater than 5. That is, if $np > 5$ and $n(1-p) > 5$. This condition is the same as required for the application of the normal approximation to the binomial probability distribution described in Chapter 3 (Section 3.5).

Example E4.12

A company manufactures engine cylinders. The machine that is used to make these cylinders is known to produce 6% defective cylinders. If a sample of 100 cylinders are selected every week and inspected them for being good or defective. If 8% or more of the cylinders in the sample are defective, the process is stopped and the machine is readjusted. Determine the probability that based on a sample of 100 cylinders the process will be stopped to readjust the machine.

SOLUTION:

Here, $p = 0.06$, $q = 1 - p = 1 - 0.06 = 0.94$ and $n = 100$.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.06(0.94)}{100}} = 0.023749$$

For $\hat{p} = 0.08$, we have

$$z = \frac{(0.08 - 0.06)}{0.023749} = 0.84$$

Therefore, $P(\hat{p} \geq 0.08) = P(z \geq 0.84) = P(1 - z \leq 0.84)$. Refer to the table in Appendix-*E*.

$P(\hat{p} \geq 0.08) = (1 - 0.799546) = 0.20045$

Example E4.13

Let m be the mean annual salary of faculty members in a college for 1995. Assume that the standard deviation of the salaries of these faculty members is \$50,000. What is the probability that the 1995 mean salary of a random sample of 100 faculty members was within \$10,000 of the population mean, m? Assume that $\frac{n}{N} \leq 0.05$.

SOLUTION:

Given $\sigma = \$50{,}000$ and $n = 100$

Standard deviation of the sample mean

$$\mu_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50000}{\sqrt{100}} = \$5000$$

The required probability is

$P[\mu - 10{,}000 \leq \bar{x} \leq \mu + 10{,}000]$

For $\bar{x} = \mu - 10{,}000$: $z = [(\mu - 10{,}000) - \mu]/5{,}000 = -2$

For $\bar{x} = \mu + 10{,}000$: $z = [(\mu + 10{,}000) - \mu]/5{,}000 = 2$

Hence $P(\mu - 10{,}000 \le \bar{x} \le \mu + 10{,}000) = P(-2 \le z \le 2)$

$= (-0.022750 + 0.977250)$ from the table in Appendix-*E*

$= 0.9545$

Example E4.14

A particular city is planning to build a nuclear power plant to generate the electric power. An independent survey fund that 53% of the voters in that city favour the building of that plant. Assume that this result holds true for the population of all the voters in this city.

(a) what is the probability that more than 50% of the voters in a random sample of 200 voters selected from this city will favour the building of this plant?

(b) a city official would like to take a random sample of voters in which over 50% would favour the plant building. How large a sample should be selected so that the city official is 95% sure of this outcome? Assume $\frac{n}{N} \le 0.05$.

SOLUTION:

Given $p = 0.53$, $n = 200$ and $n/N = \le 0.05$

(a) Standard deviation of the sample mean

$$\sigma_{\bar{x}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.53)(0.47)}{200}} = 0.03529$$

The shape of the sampling distribution is approximately normal. In order to have over 50% in favour in a sample of 200 requires 101 or more in favour of the plant building. Hence, we require

$$P\left(\hat{p} \ge \frac{101}{200}\right) = P(\hat{p} \ge 0.505)$$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.505 - 0.53}{0.03529} = -0.71$$

$$P(\hat{p} \ge 0.505) = P(z \ge -0.71) = 1 - P(z \le -0.71) = 1 - 0.238852 = 0.76115$$

(from the table in Appendix-*E*).

(b) $P(z > -1.65) = 1 - P(z < -1.65) = 1 - 0.049471 = 0.95053$ (from the table in Appendix-*E*).

Since $z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$

$$\sigma_{\hat{p}} = \frac{\hat{p} - p}{z} = \frac{0.5 - 0.53}{-1.65} = 0.01818$$

Since $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

$$n = \frac{p(1-p)}{\left(\sigma_{\hat{p}}\right)^2} = \frac{0.53(0.47)}{(0.01818)^2} = 753.53 \approx 754$$

Hence, the sample should include at least 754 voters.

Example E4.15

A sample of 49 is picked at random from a population of manufactured steel circular rods. If the standard deviation of the distribution of their diameters is known to be 3 cm, find the standard error of the mean if

(a) the population consists of 1000 steel rods

(b) the population is extremely large

SOLUTION:

(a) We are giving that $n = 49$, $N = 1000$, and $\sigma = 3$. Hence, the standard error of the mean is

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3}{\sqrt{49}}\sqrt{\frac{1000-49}{1000-1}} = \frac{3}{7}(0.97568) = 0.418$$

(b) If the population size is extremely large (practically infinite), the standard error is given by

$$\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{49}} = \frac{3}{7} = 0.4286.$$

4.10 THE CHI-SQUARE DISTRIBUTION

The *chi-square distribution* is sometimes written as χ^2 *distribution* and read as *chi-square distribution*. The values of a chi-square distribution are denoted by the symbol χ^2, just as the values of the standard normal distribution and the *t*-distribution are defined by *z* and *t*, respectively. A variable has a *chi-square distribution* if its distribution has the shape of a specific type of right-skewed curve, called a *chi-square (χ^2) curve*. There are infinitely many chi-square distributions. The chi-square curve is identified by its number of degrees of freedom.

Let $Z_1, Z_2, \ldots, Z_k$ be normally and independently distributed random variables, with mean $\mu = 0$ and variance $\sigma^2 = 1$. Then the random variable

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

has the probability density function

$$f_{\chi^2}(u) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} u^{(k/2)-1} e^{-u/2} \qquad u > 0$$

$$= 0 \qquad \text{otherwise} \tag{4.20}$$

and is said to follow the chi-square distribution with *k* degrees of freedom abbreviated χ_k^2.

The mean and variance of the χ_k^2 distribution are

$$\mu = k \tag{4.21}$$

and

$$\sigma^2 = 2k \tag{4.22}$$

Several chi-square distributions are shown in Fig. 4.1.

Some basic properties of χ^2-curves are:

1. The total area under a χ^2-curve equals 1.
2. A χ^2-curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.
3. A χ^2-curve is right-skewed.
4. As the number of degrees of freedom becomes larger, χ^2-curves look increasingly like normal curves.

Percentages (and probabilities) for a variable having a chi-square distribution are equal to areas under its associated χ^2-curve.

The chi-square random variable is non-negative, and the probability distribution is skewed to the right. As k increases, the distribution becomes more symmetric. As $k \to \infty$, the limiting form of the chi-square distribution is the normal distribution. The percentage points of the χ^2_k distribution are given in Appendix-*F*. Define $\chi^2_{\alpha,k}$ as the percentage point or value of the chi-square random variable with k degrees of freedom such that the probability that χ^2_k exceed this value is α. That

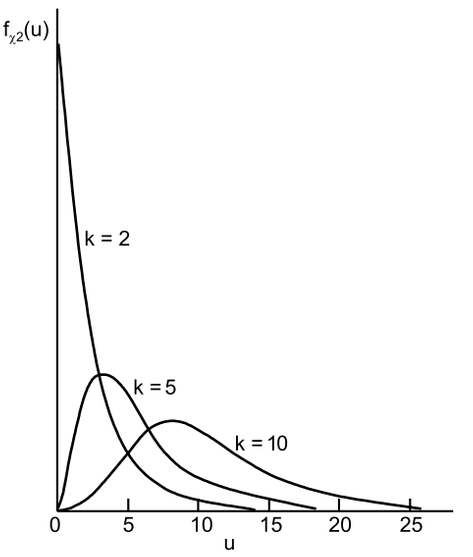

Fig. 4.1: Several χ^2 distributions

$$P\left\{\chi^2_k \geq \chi^2_{\alpha,k}\right\} = \int_{\chi^2_{\alpha,k}}^{\infty} f_{\chi^2}(\mu)\, du = \alpha \tag{4.23}$$

The probability in Eq. (4.23) is shown in Fig. 4.2.

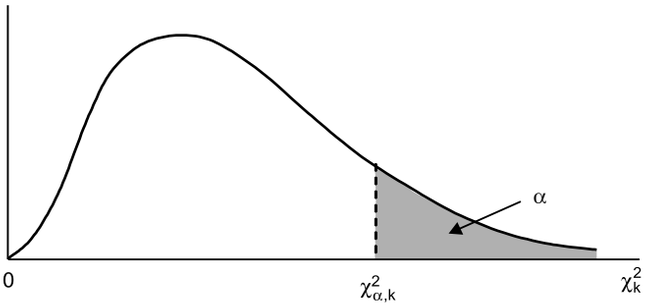

Fig. 4.2: Percentage point $\chi^2_{\alpha,k}$ of the chi-square distribution

The two outside columns of table in Appendix-*F*, labeled *df*, display the number of degrees of freedom. The symbol χ^2_α denotes the χ^2-value having area α to its right under a χ^2-value. Hence, the column headed $\chi^2_{0.995}$, for example, contains χ^2-values having area 0.995 to their right.

Example E4.16

Determine the χ^2-value having area 0.05 to the left for a χ^2-curve with $df = 9$ as shown in Fig. E4.16(*a*).

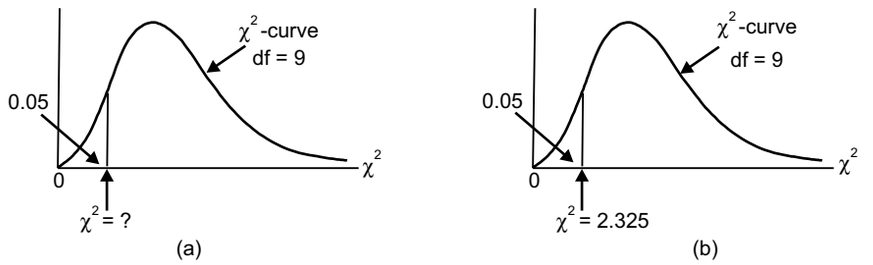

Fig. E4.16

SOLUTION:

The total area under a χ^2-curve equals $1 - 0.05 = 0.95$. Hence, the required χ^2-value is $\chi^2_{0.995}$. From the table in Appendix-*F* with $df = 9$, $\chi^2_{0.995} = 3.325$. Therefore, for a χ^2-curve with $df = 9$, the χ^2-values having area 0.05 to its left is 3.325 as shown in Fig. E4.16(*b*).

Example E4.17

For a χ^2-curve with $df = 25$, determine the two χ^2-values that divide the area under the curve into a middle 0.95 area and two outside 0.025 areas as shown in Fig. E4.17(*a*).

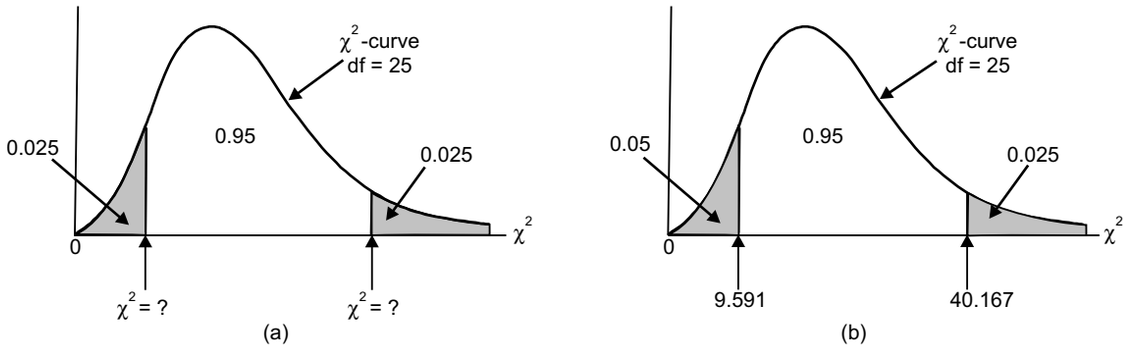

Fig. E4.17

SOLUTION:

First, we find the χ^2-value on the right in Fig. E4.17(*a*). Because the shaded area on the right is 0.025, the χ^2-value on the right is $\chi^2_{0.025}$. From the table in Appendix-*F* with $df = 25$, $\chi^2_{0.025} = 40.647$.

Next, we find the χ^2-value on the left in Fig. E4.17(*a*). Because the area to the left of that χ^2-value is 0.025, the area to its right is $1 - 0.025 = 0.975$. Hence, the χ^2-value on the left is $\chi^2_{0.975}$, which, by table in Appendix-*F* equals 13.120 for $df = 25$.

Consequently, for a χ^2-curve with $df = 25$, the two χ^2-values that the area under the curve into a middle 0.95 area and two outside 0.025 areas are 13.120 and 40.167, as shown in Fig. E4.17(*b*).

4.11 THE t-DISTRIBUTION

The standardised version of $\bar{x}$ has the standard normal distribution. William Gosset in 1908 determined the distribution of the studentised version of $\bar{x}$, a distribution now called *student's t-distribution* or, simply, the *t-distribution*.

There is a different *t*-distribution for each sample size. A particular *t*-distribution is identified by its number of *degrees of freedom* (*df*). For the studentised version of $\bar{x}$, the number of degrees of freedom is 1 less than the sample size, which is indicated symbolically by $df = n - 1$. The mean of the *t*-distribution is 0 just like the standard normal distribution. But unlike the standard normal distribution, whose standard deviation is 1, the standard deviation of a *t*-distribution is $\sqrt{\dfrac{df}{(df-2)}}$ which is always greater than 1. Hence, the standard deviation of a *t*-distribution is larger than the standard deviation of the standard normal distribution.

The number of degree of freedom is the only parameter of the *t*-distribution. A variable with a *t*-distribution has an associated curve, called a *t-curve*. There is a different *t*-distribution for each number of degrees of freedom.

The *t*-distribution is a specific type bell-shaped distribution with a height and a wider spread than the standard normal distribution. As the sample size becomes larger, the *t*-distribution approaches the standard normal distribution.

Let $Z \sim N(0, 1)$ and V be a chi-square random variable with k degrees of freedom. If Z and v are independent, then the random variable

$$\Gamma = \frac{Z}{\sqrt{V/k}} \tag{4.24}$$

has the probability density function

$$f(t) = \frac{\Gamma[(k+1)/2]}{\sqrt{\pi k}\,\Gamma(k/2)} \times \frac{1}{[(t^2/k)+1]^{(k+1)/2}} \qquad -\infty < t < \infty \tag{4.25}$$

and is said to follow the *t*-distribution with k degrees of freedom, abbreviated t_k.

The mean and variance of t are $\mu = 0$ and $\sigma^2 = k/(k-2)$ for $k > 2$, respectively. Several *t*-distributions are shown in Fig. 4.3.

The basic properties of *t*-curves are:

1. The total area under a *t*-curve equals 1.
2. A *t*-curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.
3. A *t*-curve is symmetric about 0.
4. As the number of degrees of freedom becomes larger, *t*-curves look increasingly like the standard normal curve.

Note that if $k = \infty$, the t-distribution becomes the standard normal distribution. A table of percentage points of the t-distribution is given in Appendix-G. If $x_1, x_2, \ldots, x_n$ is a random sample from $N(\mu, \sigma^2)$ distribution, then the quantity

$$t = \frac{x - \mu}{s/\sqrt{n}}$$

is distributed as t with $n - 1$ degrees of freedom.

We will let $t_{\alpha,k}$ be the value of the random variable Γ with k degrees of freedom above which we find an area (or probability) α. Thus $t_{\alpha,k}$ is an upper-tail 100α percentage point of the t-distribution with k degrees of freedom. This percentage point is shown in Fig. 4.4. In the Appendix-G, the α values are the column headings, and the degrees of freedom are listed in the left column. To illustrate the use of the table, note that the t-value with 15 degrees of freedom having an area of 0.05 to the right is $t_{0.05,10} = 1.753$. That is,

$$P(\Gamma_{10} > t_{0.05,15}) = P(\Gamma_{10} > 1.753) = 0.05$$

Since t-distribution is symmetric about zero, we have $t_{1-\alpha} = -t_{\alpha}$; that is, the t-value having an area of $1 - \alpha$ to the right (and therefore an area of α to the left) is equal to the negative of the t-value that has area α in the right tail of the distribution. Therefore, $t_{0.95,15} = -t_{0.05,15} = -1.753$.

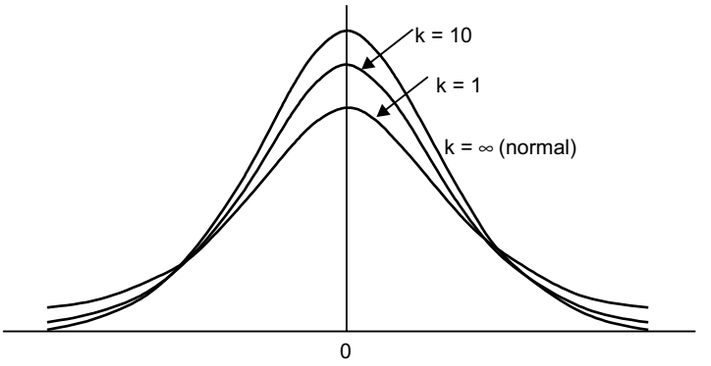

Fig. 4.3: Several *t*-distributions

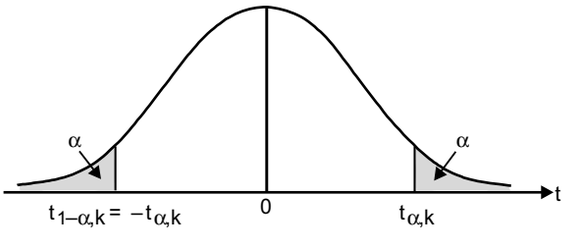

Fig. 4.4: Percentage points of the *t*-distribution

Example E4.18

Find a t-curve with 17 degrees of freedom, determine $t_{0.05}$. In other words, find the t-value having area 0.05 to its right, as shown in Fig. E4.18(a).

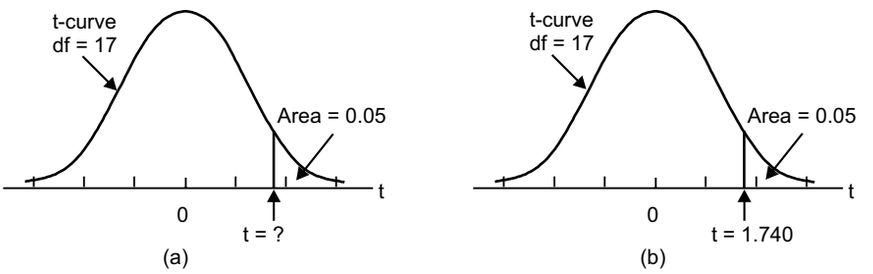

Fig. E4.18

SOLUTION:

The number degrees of freedom is 17, so we first go down the outside columns, (from the table in Appendix-*E*) labeled *df*, to "17". Then, going across that row to the column labeled $t_{0.05}$, we reach 1.740. This number is the *t*-value having area 0.05 to its right, as shown in Fig. E4.18(*b*). In other words, for a *t*-curve with $df = 17$, $t_{0.05} = 1.740$.

4.12 THE F-DISTRIBUTION

A variable is said to have an *F-distribution* if its distribution has the shape of a special type of right-skewed curve, called an *F-curve*. There are infinitely many *F*-distributions. The shape of a particular *F-distribution* curve depends on the number of degrees of freedom. There are two numbers of degrees of freedom for the *F*-distribution curve. The first number of degrees of freedom for an *F*-curve is called the *degrees of freedom for the numerator* and the second the *degrees of freedom for the denominator.*

The random variable *F* is defined to be the ratio of two independent chi-square random variables, each divided by its number of degrees freedom. That is

$$F = \frac{W/u}{Y/v} \tag{4.26}$$

where *W* and *Y* are independent chi-square random variables with *u* and *v* degrees of freedom, respectively.

Let *W* and *Y* be independent chi-square random variables with *u* and *v* degrees of freedom, respectively. Then the ratio

$$F = \frac{W/u}{Y/v} \tag{4.27}$$

has the probability density function

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}} \qquad 0 < x < \infty \tag{4.28}$$

and is said to follow the *F*-distribution with *u* degrees of freedom in the numerator and *v* degrees of freedom in the denominator. It is usually abbreviated as $F_{u,v}$.

The mean and variance of *F*-distribution are $\mu = v/(v-2)$ for $v > 2$, and

$$\sigma^2 = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)} \qquad v > 4$$

The F random variable is non-negative, and the distribution is skewed to the right. The F-distribution looks very similar to the chi-square distribution as shown in Fig. 4.5; however the two parameters u and v provide extra flexibility regarding shape.

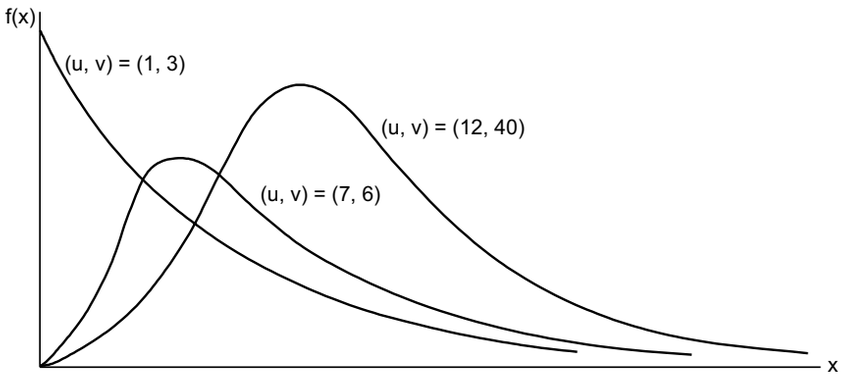

Fig. 4.5: Three *F*-distribution curves

The basic properties of F-curves are:

1. The total area under an F-curve equals 1.
2. An F-curve starts on 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.
3. An F-curve is right-skewed.

For an F-curve with $df = (u, v)$, the F-value having area α to its left equals the reciprocal of the F-value having area α to its right for an F-curve with $df = (v, u)$.

The percentage points of F distribution are given in Appendix-*H*. Let $f_{\alpha,u,v}$ be the percentage point of F-distribution, with the numerator degrees of freedom u and the denominator degrees of freedom v such that the probability that the random variable F exceeds this value is

$$P(F > f_{\alpha,u,v}) = \int_{f_{\alpha,u,v}}^{\infty} f(x)\,dx = \alpha \tag{4.29}$$

This is illustrated in Fig. 4.6.

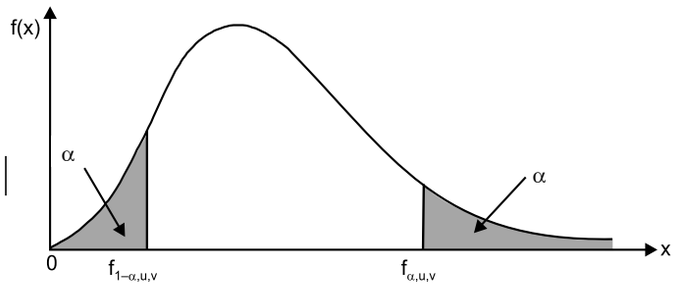

Fig. 4.6: Upper and lower percentage points of the F-distribution

For example, if $u = 6$ and $v = 10$, we find from the table in Appendix-*H*, that

$$P(F > f_{0.05,5,10}) = P(F_{5,10} > 3.22) = 0.05$$

That is, the upper 5 percentage point of $F_{5,10}$ is $f_{0.05,5,10} = 3.22$. The table in Appendix-*H* contains only upper-tail percentage points (for selected values of $f_{\alpha,u,v}$ for $\alpha \leq 0.25$) of the *F*-distribution. The lower-tail percentage point $f_{1-\alpha,u,v}$ can be found as follows:

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,u,v}} \tag{4.30}$$

For example, to find the lower-tail percentage point of $f_{0.95,6,10}$, note that

$$f_{0.95,6,10} = \frac{1}{f_{0.05,10,6}} = \frac{1}{4.06} = 0.246$$

Example E4.19

For an *F*-curve with *df*(5, 13), find $F_{0.05}$. That is, find an *F*-value having area 0.05 to its right as shown in Fig. E4.19(*a*).

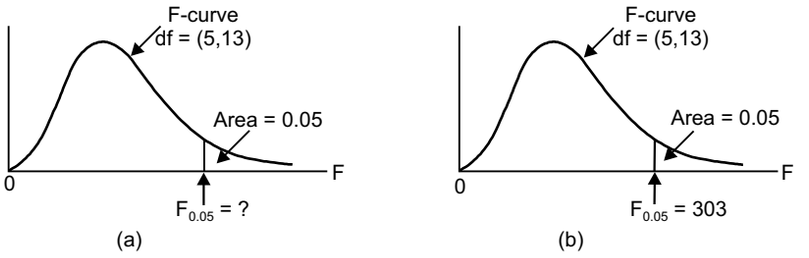

Fig. E4.19

SOLUTION:

To find the *F*-value, we use the table in Appendix-*H*. In this case, $\alpha = 0.05$ (area in the right tail under the *F*-distribution curve). The degrees of freedom for the numerator is 5, and the degrees of freedom for the denominator is 13.

We first go down the degree of freedom to "13". Next, we concentrate on the row for α labeled 0.05. Then, going across that row to the column labeled "5" (from the table in Appendix-*E*), we reach 3.03. This number is the *F*-value having area 0.05 to its right, as shown in Fig. E4.19(*b*). In other words, for an *F*-curve with *df*(5, 15), $F_{0.05} = 3.03$.

Example E4.20

For an *F*-curve with *df* = (50, 8) having an area 0.05 to its left.

SOLUTION:

The required *F*-value is the reciprocal of the *F*-value having area 0.05 to its right for an *f*-curve with *df* = (8, 50). From the table in Appendix-*H*, this latter *F*-value equals 2.13. Consequently, the required *F*-value is $\frac{1}{2.13}$ or 0.4695, as shown in Fig. E4.20.

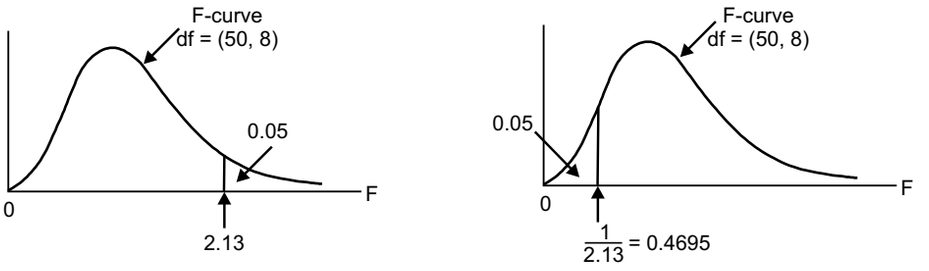

Fig. E4.20

Example E4.21

For an F-curve with $df = (10, 8)$, determine the two F-values that divide the area under the curve into a middle 0.95 area and two outside 0.025 areas as shown in Fig. E4.21(a).

SOLUTION:

First, we find the F-value on the right in Fig. E4.21(a). Because, the shaded area on the right is 0.025, the F-value on the right is $F_{0.025}$. From the table in Appendix-H, with $df = (10, 8)$, $F_{0.025} = 4.30$.

Next, we find the F-value on the left in Fig. E4.21(a). The F-value is the reciprocal of the F-value having area 0.025 to its right for an F-curve with $df(10, 8)$. From the table in Appendix-H, we find that this latter F-value equals $\dfrac{1}{4.30} = 0.2326$.

Consequently, for an F-curve with $df(10, 8)$, the two F-values that divide the area under the curve into a middle 0.95 area and two outside 0.025 areas are 0.2326 and 4.30, as shown in Fig. E4.21(b).

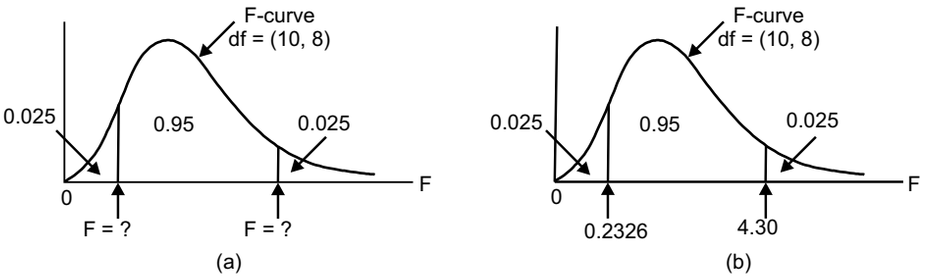

Fig. E4.21

4.13 SUMMARY

In Chapter 3, we discussed the probability distributions of discrete and continuous random variables. In this chapter, we extended the concept of probability distribution to that of a *sample statistic*. A sample statistic is a numerical summary measure calculated for sample data. The numerical summary measures calculated for population data are called *population parameters*. A population parameter is always a constant, whereas a sample statistic is always a random variable. Also, since every random variable must possess a probability distribution, each sample statistic possesses a probability distribution. The probability distribution of a sample statistic is more generally called its *sampling distribution*. In this chapter, we presented the sampling distributions of the sample mean and the sample proportion. The concepts presented in this chapter forms the foundation for the inferential statistics discussed in Chapters 5 and 6.

PROBLEMS

P4.1 A population consists of the values 1, 2 and 5.

(a) list all the possible samples (with replacement) of size $n = 2$ along with the sample means and their individual probabilities.

(b) find the mean of the sampling distribution

(c) does the sample mean target the values of the population mean?

P4.2 A manufacturing company produces machine parts that have a mean tensile strength of 100 MPa and a standard deviation of 10 MPa. The distribution of tensile strength is normal. Find the probability that a tensile strength less than 95 MPa.

P4.3 Suppose that a random variable X has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2, & 4 \le x \le 6 \\ 0, & \text{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 50$.

P4.4 Assume that the weights of 4000 male students at a university are normally distributed with a mean of 60 kg and standard deviation 3.0 kg. If 100 samples consisting 25 students each are obtained, what would be the expected mean and standard deviation of the resulting sampling distribution of means if sampling were done (*a*) with replacement, (*b*) without replacement?

P4.5 A survey showed that the mean expenditure incurred by a student in 1995 was \$9000 and the standard deviation of the expenditure was \$800. Find the approximate probability that the mean expenditure of 64 students picked at random was

(a) more than \$8820

(b) between \$8800 and \$9120

P4.6 Refer to Problem P4.4. In how many samples of Problem P4. 4 would you expect to find the mean

(a) between 58.8 and 60.3 kg

(b) less than 58.4 kg

P4.7 The length of life (in hours) of a certain type of machine component is a random variable with a mean of 600 hours and a standard deviation of 70 hours. What is the approximate probability that a random sample of 196 machine parts will have a mean life between 588 and 605 hours?

P4.8 Five hundred electronic components have a mean weight of 6.02 g and a standard deviation of 0.3 g. Find the probability that a random sample of 100 electronic components chosen from this group will have a combined weight of

(a) between 596 and 600 g

(b) more than 610 g

P4.9 Assume that the weights of machine parts of a heavy machine are normally distributed with mean 55 kg and standard deviation 2 kg. A number of samples of 100 parts each are taken at random with replacement from the population. Determine

(a) the mean and standard deviation of the sampling distribution of the mean

(b) the probability that the sample mean will differ from the population mean by less than 0.4 kg.

P4.10 It was found in a particular survey that adults spend an average of 10 hours a day at work and commuting. Let the daily work and commute times for all adults have a mean of 10 hours and a standard deviation of 2 hours. Find the probability that the mean of the daily work and commute times for a random sample of 100 adults will be

(a) greater than 10.45 hours

(b) between 9.75 and 10.50 hours

(c) within 0.25 hours of the population mean

(d) lower than the population mean by 0.50 hours or more.

P4.11 A random sample of size 121 is taken from a normal population with $\sigma = 27.5$. What is the probability that the mean of the sample will differ from the mean of the population by 3 or more either way?

P4.12 The diameters of ball bearings manufactured by a company are normally distributed with mean 12 mm and standard deviation 0.1 mm. A sample of 25 ball bearings was taken each day during a given month. Find

(a) the measured standard deviation of the distribution of the sample means

(b) the probability that the sample mean of ball bearings will

(i) exceed 12.01 mm

(ii) be less than 11.98 mm

(iii) lie between 11.98 and 12.1 mm

P4.13 A random sample of size of 81 is taken from a normal population with $\mu = 61.4$ and $\sigma = 7.65$. What is the probability that the mean of the sample will

(a) exceed 62.9

(b) fall between 60.5 and 62.3

(c) be less than 60.6

P4.14 A random sample of 300 air compressors had a standard deviation of 50 months. Determine the probability that the estimate on the population mean lifetime of these compressors will be within 6 months of the true mean from the sample estimate.

P4.15 According a particular survey, the standard deviation of the lengths of time that men with one job are employed during the first 10 years of their career is 100 weeks. Length of time employed during the first 10 years of career is left-skewed variable. For this variable, find the following:

(a) the sampling distribution of the sample mean for simple random samples of 100 men with one job

(b) the probability that the sampling error made in estimating the mean length of time employed by all men with one job by that of a random sample of 100 such men will be at most 25 weeks.

P4.16 In a large university, it was found that the score in statistics class with large number students over a number of years has a distribution with mean 80 and standard deviation 16. A random sample of 64 students of every class for 20 classes is taken. For the sampling distribution of mean, find

(a) the mean and standard deviation

(b) the probability that

(i) $\bar{X} > 84$

(ii) $\bar{X} > 86$

(iii) $\bar{X} \leq 80$

where $\bar{X}$ is the sample mean found for the sample size of 64.

P4.17 According to a particular survey, the mean annual salary of private classroom teachers is \$45,000 and standard deviation of \$9000.

(a) find the sampling distribution of the sample mean for sample size of 81

(b) repeat part *(a)* for samples of 289

(c) is the assumption that classroom teachers salaries are normally distributed to answers in parts *(a)* and *(b)* necessary?

(d) what is the probability that the sampling error made in estimating the population mean salary of all classroom teachers by the mean salary of a sample of 81 classroom teachers will be at most \$1000

(e) repeat part *(d)* for samples of size 289.

P4.18 The weight of electronic components packed in certain containers is a random variable with a mean weight of 16 g and a standard deviation of 0.6 g. If the containers are shipped in boxes of 36, find, approximately, the probability that a randomly picked box will weigh over 585 g.

P4.19 According a particular survey it was found that the standard deviation of the lengths of hospital stay on the intervention ward is 9 days.

(a) for the variable "length of hospital stay", determine the sampling distribution of the sample mean for samples of 100 patients on the intervention ward

(b) the distribution of the length of hospital stay is right-skewed. Does this validate or invalidate the results found in part *(a)*?

(c) find the probability that the sampling error made in estimating the population mean length of stay on the intervention ward by the mean length of stay of a sample of 100 patients will be at most 3 days.

P4.20 The proportion of employed men between the ages 21 and 40 years in a city is 2/3. Suppose random samples of size 25 are drawn with replacement from all men in that city between the ages 21 and 40. What are the mean and standard deviation of the proportion $\hat{p}$ for all such samples?

P4.21 In a particular survey in a large metropolitan city nearly 20% of all students were at least somewhat afraid of being attacked in or nearby their schools. Suppose 20% of all students in grade 10 are afraid of such attacks. If $\hat{p}$ is the proportion of students in a random sample of 49 of grade 10 students who fear such attacks, find the mean and standard deviation of $\hat{p}$.

P4.22 Refer to Problem P4.20. Suppose the city has 250 men between ages 21 and 40 years, and the sampling is without replacement. What are the mean and standard deviation of $\hat{p}$?

P4.23 Suppose 25% of all workers in a state belong to a labour union. What is the probability that in a random sample of 100 workers in that state, at least 18% will belong to a labour union?

REVIEW QUESTIONS

1. Explain the meaning of a population distribution and a sampling distribution.
2. What is sampling error? Does such an error occur only in a sample survey or can it occur both in a sample survey and a census?
3. Explain the meaning of nonsampling error. Do such errors occur only in a sample survey or can they occur both in a sample survey and a census?
4. Describe the condition or conditions that must hold true for the sampling distribution of the sample mean to be normal when the sample size is less than 30.
5. Describe the central limit theorem (Refer to Chapter 3, Section 3.8).
6. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 2.5 standard deviations of the population mean? (Chapter 3, Section 3.9).
7. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 1.5 standard deviations of the population mean? (Chapter 3, Section 3.9).
8. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 2.0 standard deviations of the population mean? (Chapter 3, Section 3.9).
9. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 3.0 standard deviations of the population mean? (Chapter 3, Section 3.9).
10. Define the following terms:
    (a) Consistent estimator
    (b) Estimator
    (c) Mean of $\hat{p}$
    (d) Mean of $\overline{x}$
    (e) Population proportion, p
    (f) Unbiased estimator
11. Explain the following terms:
    (a) Sample proportion
    (b) Sampling distribution of $\hat{p}$
    (c) Sampling distribution of $\overline{x}$
    (d) Standard deviation of $\hat{p}$
    (e) Standard deviation of $\overline{x}$
12. Describe very briefly the following distributions:
    (a) t-distribution
    (b) F-distribution
    (c) chi-square distribution

STATE TRUE OR FALSE

1. Sampling error is the error resulting from using a sample to estimate a population characteristic. (True/False)
2. For a variable x and a given sample size, the distribution of the variable $\bar{x}$ is called the sampling distribution of the population mean. (True/False)
3. The larger the sample size, the smaller the sampling error tends to be in estimating a population mean, μ, by a sample mean $\bar{x}$. (True/False)
4. For a sample of size n, the mean of the variable $\bar{x}$ equals the mean of the variable under consideration. (True/False)
5. For samples of size n, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration multiplied by the square root of the sample size. (True/False)
6. A normal distribution is determined by the mean and standard deviation. (True/False)
7. The mean of all possible sample means (i.e., of the variable $\bar{x}$) always equals the population mean. (True/False)
8. The larger the sample size, the larger is the standard deviation $\bar{x}$. (True/False)
9. The smaller the standard deviation of $\bar{x}$, the more closely the possible values $\bar{x}$ of (the possible sample means) cluster around the mean of $\bar{x}$. (True/False)
10. For a relatively large sample size, the variable $\bar{x}$ is approximately normally distributed.(True/False)
11. The total area under χ^2-curve equals 1. (True/False)
12. A χ^2-curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so. (True/False)
13. A χ^2-curve is left skewed. (True/False)
14. As the number of degrees of freedom becomes larger, χ^2-curves look decreasingly like normal curves. (True/False)
15. The larger the standard deviation of $\bar{x}$, the more closely the possible values of $\bar{x}$ (the possible sample means) cluster around the mean of $\bar{x}$. (True/False)
16. For samples of size n, the variable $\chi^2 = \left(\frac{n-1}{\sigma^2}\right)s^2$ has the chi-square distribution with $n - 1$ degrees of freedom. (True/False)
17. A variable is said to have a chi-square distribution if its distribution has the shape of a special type of right-skewed curve, called a chi-square curve. (True/False)
18. Different chi-square distributions are identified by their degrees of freedom. (True/False)
19. The total area under an F-curve equals 1. (True/False)
20. An F-curve starts at 0, on the horizontal axis and extends indefinitely to the left, approaching, but never touching, the horizontal axis as it does so. (True/False)
21. An F-curve with $df = (v_1, v_2)$, the F-value having area α to its left equals the reciprocal of the F-value having area α to its right for an F-curve with $df = (v_2, v_1)$. (True/False)
22. An F-distribution and its corresponding F-curve are identified by stating its two numbers of degrees of freedom. (True/False)
23. An F-distribution has two numbers of degrees of freedom. (True/False)

24. The first degree of freedom in F-distribution is called the degrees of freedom for the numerator. (True/False)
25. The second degree of freedom in F-distribution is called the degrees of freedom for the denominator. (True/False)
26. A χ^2-curve looks increasingly like a normal curve as the number of degrees of freedom becomes larger. (True/False)
27. An F-curve is right-skewed. (True/False)
28. For an F-curve with $df = (15, 5)$, the F-value having area 0.05 to its left equals the reciprocal of the F-value having area 0.05 to its right for an F-curve with $df = (5, 15)$. (True/False)
29. The observed value of a variable having an F-distribution must be greater than or equal to 1. (True/False)
30. The total area under a t-curve equals 1. (True/False)
31. A t-curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so. (True/False)
32. A t-curve is not symmetric about 0. (True/False)
33. As a number of degrees of freedom becomes larger, t-curves look increasingly like the F-distribution curve. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **True** 5. **False** 6. **True** 7. **True** 8. **False** 9. **True** 10. **True**
11. **True** 12. **True** 13. **False** 14. **False** 15. **False** 16. **True** 17. **True** 18. **True** 19. **True** 20. **False**
21. **True** 22. **True** 23. **True** 24. **True** 25. **True** 26. **True** 27. **True** 28. **True** 29. **False** 30. **True**
31. **True** 32. **False** 33. **False**