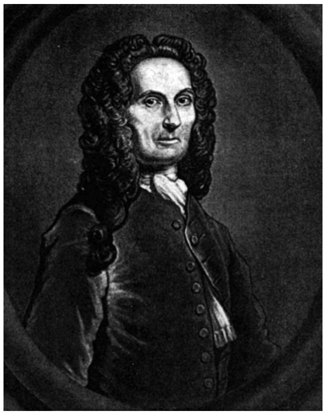
Chapter 4

Sampling distributions

Chapter outline

Objective

In this chapter we study the probability distributions of various sample statistics such as the sample mean and the sample variance and illustrate their usefulness.



Abraham de Moivre
(Source: http://en.wikipedia.org/wiki/File:Abraham_de_Moivre.jpg.)

Abraham de Moivre (1667—1754) was a French mathematician known for his work on normal distribution and probability theory. He is famous for de Moivre’s formula, which links complex numbers and trigonometry. He fled France and went to England to escape the persecution of Protestants. In England he wrote a book on probability theory, titled *The Doctrine of Chances*. This book was very popular among gamblers. The normal distribution was first introduced by de Moivre in an article in 1733 in the context of approximating certain binomial distributions for large *n*, and this approximation result is now called the theorem of de Moivre—Laplace.

Mathematical Statistics with Applications in R. https://doi.org/10.1016/B978-0-12-817815-7.00004-X

4.1 Introduction

Sampling probability distributions plays a very important role in statistical analysis and decision-making. We begin with studying the distribution of a statistic computed from a random sample. Based on the probabilistic foundation of Chapters 2 and 3, the present study marks the beginning of our learning of statistics beyond the descriptive phase. Because a sample is a set of random variables, $X_1, \ldots, X_n$, it follows that a sample statistic that is a function of the sample is also random. We call the probability distribution of a sample statistic its *sampling distribution*. Sampling distributions provide the link between probability theory and statistical inference. The ability to determine the distribution of a statistic is a critical part in the construction and evaluation of statistical procedures. It is important to observe that there is a difference between the distribution of the population from which the sample was taken and the distribution of the sample statistic. In general, a population has a distribution called a population distribution, which is usually unknown, whereas a statistic has a sampling distribution, which is usually different from the population probability distribution. *The sampling distribution of a statistic provides a theoretical model of the relative frequency histogram for the likely values of the statistic that one would observe through repeated sampling*. Even though some of the terms in this section have already been defined in Chapter 1, we now present these definitions in terms of random variables. These abstractions are introduced to develop scientifically based methods of analyzing the data, and one should always keep in mind the underlying population.

Definition 4.1.1 *A* **sample** *is a set of observable random variables,* $\mathrm{X_1, \ldots, X_n}$. *The number* n *is called the* **sample size**.

In most of the inferential procedures that we study in this book, we are dealing with random samples. We call the random variables $X_1, \ldots, X_n$ *identically distributed* if every X_i has the same probability distribution.

Definition 4.1.2 *A* **random sample** *of size* n *from a population is a set of* n *independent and identically distributed* (*iid*) *observable random variables* $\mathrm{X_1, \ldots, X_n}$.

Note that in a sample (not a random sample), X_i need not be independent or identically distributed. For the results in this book to be applicable, it is important to ensure that the selection of a sample is at least approximately random. The significance of random sampling is that the probability distribution of a statistic can be easily derived. Random sampling helps us to control systematic biases. For a finite population, one can serially number the elements of the population and then select a random sample with the help of a table of random digits. One of the simplest ways to select a random sample of finite size is to use a table of random numbers. When the population size is very large, such a method can become very taxing and sometimes practically impossible. However, there are excellent computer programs for generating random samples from large populations, and these programs can be used. Now we define a statistic.

Definition 4.1.3 *A function* T *of observable random variables* $\mathrm{X_1, \ldots, X_n}$ *that does not depend on any unknown parameters is called a* **statistic.**

The sample mean $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$ is a function of $X_1, \ldots, X_n$. The sample median and sample variance S^2 are also examples of statistics. It is important to observe that even with random sampling, there is sampling variability or error. That is, if we select different samples from the same population, a statistic will take different values in different samples. Thus, a sample statistic is a random variable, and hence it has a probability distribution. For us to study the behavior of the phenomenon a sample statistic represents, we must identify its probability distribution.

Definition 4.1.4 *The probability distribution of a sample statistic is called the* **sampling distribution.**

We can illustrate these definitions with the following example with a finite population and a finite sample size. In this case, we take all possible samples of size n from a population of size N.

EXAMPLE 4.1.1

Let the population consist of the numbers {1, 2, 3, 4, 5}. Consider all possible samples consisting of three numbers randomly chosen without replacement from this population. Obtain the distribution of the sample mean.

Solution

Disregarding the order, it is clear that there are $\binom{5}{3} = 10$ *equally likely possible samples of size 3. They are (1, 2, 3), (1, 2, 4), (1, 2, 5), (1, 3, 4), (1, 3, 5), (1, 4, 5), (2, 3, 4), (2, 3, 5), (2, 4, 5), and (3, 4, 5). Calculating the sample mean,* $\overline{\mathrm{X}}$, *for each of the samples, we will get the sampling distribution of* $\overline{\mathrm{X}}$ *as:*

$\overline{x}$	$\frac{2}{1}$	$\frac{7}{3}$	$\frac{8}{3}$	$\frac{3}{1}$	$\frac{10}{3}$	$\frac{11}{3}$	$\frac{4}{1}$
$p(\overline{x})$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

For example, in the table, $P(\overline{X} = 8/3) = 2/10$ *because the two samples (1, 2, 5) and (1, 3, 4) both give an* $\overline{x} = 8/3$, *which is an estimate of the population mean,* μ.

In general, sampling distributions are theoretical distributions that consist of possibly an infinite number of sample statistics taken from an infinite number of randomly selected samples of a fixed sample size. For example, if a sample of size $n = 30$ were taken from a large population an infinite number of times, the combined means taken from all the samples would make up the sampling distribution of the mean. Every sample statistic has a sampling distribution. The next result states that if one selects a random sample from a population with mean μ and variance σ^2, then regardless of the form of the population distribution, one can obtain the mean and standard deviation of the statistic $\overline{X}$ in terms of the mean and standard deviation of the population. This is explained in the following result.

Theorem 4.1.1 *Let* $X_1, \ldots, X_n$ *be a random sample of size* n *from a population with mean* μ *and variance* σ^2. *Then* $E(\overline{X}) = \mu$ *and* $\mathrm{Var}(\overline{X}) = \sigma^2/n$.

Proof. The mean and variance of $\overline{X}$ are given by,

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu,$$

and

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)\left(\text{besause } X_i' \text{ s are independent and } Var(aX_i) = a^2 Var(X_i)\right)$$

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

We denote $E(\overline{X}) = \mu_{\overline{X}}$ and $Var(\overline{X}) = \sigma^2_{\overline{X}}$. Note that from the previous theorem, $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. Here, $\sigma_{\overline{X}}$ is called the *standard error* of the mean. It is important to notice that the variance of each of the random variables X_1, X_2, …, X_n is σ^2, whereas the variance of the sample mean $\overline{X}$ is σ^2/n, which is smaller than the population variance σ^2 for $n \geq 2$.

The implication of Theorem 4.1.1 is that the sample means become more and more reliable as an estimate of μ as the sample size is increased, as we would expect. From Chebyshev's inequality,

$$P\left(|\overline{X} - \mu_{\overline{X}}| < k\sigma_{\overline{X}}\right) \geq 1 - \frac{1}{k^2}.$$

Let $\varepsilon = (k\sigma/\sqrt{n})$ Then $k = (\varepsilon\sqrt{n})/\sigma$. Since $\mu_{\overline{X}} = \mu$, the above inequality can be written as

$$P\left(|\overline{X} - \mu| < \varepsilon\right) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

Thus, for any $\varepsilon > 0$, the probability that the difference between $\overline{X}$ and μ is less than ε can be made arbitrarily close to 1 by choosing sample size n that is sufficiently large. We illustrate this result in the following example.

EXAMPLE 4.1.2

A particular brand of drink is packaged at an average of 12 oz per bottle. As a result of randomness, there will be small variations in how much liquid each bottle really contains. It has been observed that the amount of liquid in these bottles is normally distributed with $\sigma = 0.8$ oz. A sample of 10 bottles of this brand of soda is randomly selected from a large lot of bottles, and the amount of liquid, in ounces, is measured in each. Find the probability that the sample mean will be within 0.5 oz of 12 oz.

Solution

Let $X_1, X_2, \ldots, X_{10}$ *denote the ounces of liquid measured for each of the bottles. We know that* X_i*s are normally distributed with mean* $\mu = 12$ *and variance* $\sigma^2 = 0.64$*. From Theorem 4.1.1,* $\overline{X}$ *possesses a normal distribution (actually, for the normality part, we use Corollary 4.2.2) with a mean 12 and variance* $\sigma^2/n = 0.64/10 = 0.064$*. We find that:*

$$\begin{aligned} P\left(\left|\overline{X} - 12\right| \le 0.5\right) &= P\left(-0.5 \le \left(\overline{X} - 12\right) \le 0.5\right) \\ &= P\left(-\frac{0.5}{\sigma/\sqrt{n}} \le \frac{\overline{X} - 12}{\sigma/\sqrt{n}} \le \frac{0.5}{\sigma/\sqrt{n}}\right) \\ &= P\left(-\frac{0.5}{0.253} \le Z \le \frac{0.5}{0.253}\right) \\ &= P(-1.97 \le Z \le 1.97) \\ &= 0.9512 \text{ (using a standard normal table).} \end{aligned}$$

Hence, the chance is about 0.95% that the mean amount of drink in any 10 bottles randomly chosen will be between 11.5 and 12.5 oz.

4.1.1 Finite populations

Let $\{c_1, c_2, \ldots, c_N\}$ be a finite population. Then the population mean $\mu = (1/N)\sum_{i=1}^{N} c_i$ and the population variance $\sigma^2 = (1/N)\sum_{i=1}^{N}(c_i - \mu)^2$. The following theorem for the sample mean and variance is stated without proof.

Theorem 4.1.2 *If* $X_1, \ldots, X_n$ *is a sample of size* n *(chosen without replacement) from a population* $\{c_1, c_2, \ldots, c_N\}$*, then:*

$$E\left(\overline{X}\right) = \mu$$

and

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

We remark here that the sample in the theorem is not a random sample and X_i' are not id random variables. The factor $(N-n)/(N-1)$ in the foregoing theorem is often called the ***finite population correction factor.*** It is close to 1 unless the sample amounts to a significant portion of the population. Note that the sampling without replacement causes dependence among the X_is. However, if the sample size n is small relative to the population size N, the population correction factor is approximately 1. Hence, we will not use the finite population correlation factor in the derivation of a sampling distribution, unless it is absolutely necessary.

EXAMPLE 4.1.3

Obtain the mean and variance of $\overline{X}$ in Example 4.1.1.

Solution

First note that for the population in Example 4.1.1, *the population mean is* $\mu = (1/N)\sum_{i=1}^{N} c_i = 3$ *and the population variance is* $\sigma^2 = (1/N)\sum_{i=1}^{N}(c_i - \mu)^2 = 2$. *Applying the probability distribution of* $\overline{X}$ *given in* Example 4.3.1, *we obtain:*

$$\begin{aligned} E\left(\overline{X}\right) &= 2\left(\frac{1}{10}\right) + \frac{7}{3}\left(\frac{1}{10}\right) + \frac{8}{3}\left(\frac{2}{10}\right) + 3\left(\frac{2}{10}\right) + \frac{10}{3}\left(\frac{2}{10}\right) + \frac{11}{3}\left(\frac{1}{10}\right) + 4\left(\frac{1}{10}\right) \\ &= 3, \end{aligned}$$

and

$$Var(\overline{X}) = E(\overline{X}^2) - E(\overline{X}^2) = 2^2\left(\frac{1}{10}\right) + \left(\frac{7}{3}\right)^2\left(\frac{1}{10}\right) + \left(\frac{8}{3}\right)^2\left(\frac{2}{10}\right)$$

$$+3^2\left(\frac{2}{10}\right) + \left(\frac{10}{3}\right)^2\left(\frac{2}{10}\right) + \left(\frac{11}{3}\right)^2\left(\frac{1}{10}\right) + 4^2\left(\frac{1}{10}\right) - 3^2$$

$$= \frac{2}{3} \times \frac{1}{2} = 0.3333.$$

This is the same as $\frac{\sigma^2}{n}\left[\frac{(N-n)}{(N-1)}\right]$. *In this case we observe that the variance of* $\overline{X}$ *is precisely one-sixth of the original variance.*

EXAMPLE 4.1.4

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. Consider the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Show that $E(S^2) = \sigma^2$.

Solution

It can be shown that (see Exercise 1.5.8):

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}.$$

Hence,

$$E(S^2) = E\left(\frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}\right) = \frac{1}{n-1}\sum_{i=1}^{n}E(X_i^2) - \frac{n}{n-1}E(\overline{X}^2).$$

Using the fact that $E(X^2) = Var(X) + \mu^2$ *and* Theorem 4.1.1, *we have*:

$$E(S^2) = \frac{1}{n-1}n(\sigma^2 + \mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu^2\right)$$

$$= \left(\frac{n}{n-1} - \frac{1}{n-1}\right)\sigma^2 + \left(\frac{n}{n-1} - \frac{n}{n-1}\right)\mu^2$$

$$= \sigma^2.$$

This shows that the expected value of the sample variance is the same as the variance of the population under consideration.

Exercises 4.1

4.1.1. Let the population be given by the numbers $\{-2, -1, 0, 1, 2\}$. Take all random samples of size 3.

(a) Without replacement, obtain the following in each case.

(i) The sampling distribution of the sample mean.

(ii) The sampling distribution of the sample median.

(iii) The sampling distribution of the sample standard deviation.

(iv) The mean and variance of the sample mean.

(b) How many samples of size 3 can we get, if we sample with replacement?

4.1.2. **(a)** How many different samples of size $n = 2$ can be chosen from a finite population of size 12 if the sampling is without replacement?

(b) What is the probability of each sample in part (a), if each sample of size 2 is equally likely?

(c) Find the value of the finite population correction factor.

4.1.3. Let the population be given by {1, 2, 3}. Let P(x) = 1/3 for $x = 1, 2, 3$. Take samples of size 3 with replacement.

(a) Calculate μ and σ^2.

(b) Obtain the sampling distribution of the sample mean.

(c) Obtain the mean and variance of the sample mean.

4.1.4. Find the value of the finite population correlation factor for

(a) $n = 8$ and $N = 60$.

(b) $n = 8$ and $N = 1000$.

(c) $n = 15$ and $N = 60$.

4.1.5. For a random sample $X_1, \ldots, X_n$, let $(S')^2 = (1/n)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Find E[(S′)^2]. Compare this with E(S^2).

4.1.6. For a random sample X$_1$, …, X_n with mean μ and variance σ^2, let $T_n = \sum_{i-1}^{n} X_i$, the sample total. Show that $E(T_n) = n\mu$ and $Var(T_n) = n\sigma^2$.

4.1.7. A particular brand of sugar is sold in 5-lb packages. The weight of sugar in these packages can be assumed to be normally distributed with mean $\mu = 5$ lb and standard deviation $\sigma = 0.2$ lb. What is the probability that the mean weight of sugar in 15 randomly selected packages will be within 0.2 lb of 5 lb?

4.1.8. A random sample of size 50 is taken from an infinite population having the mean $\mu = 15$ and standard deviation $\sigma = 4$. What is the probability that $\overline{X}$ will be between 13.5 and 16.6?

4.1.9. The distribution of heights of all students in a large university has a normal distribution with a mean of 66 in. and a standard deviation of 2 in. What is the probability that the mean height of 26 randomly selected students from this university will be more than 67 in.?

4.1.10. An image-encoding algorithm, when used to encode images of a certain size, uses a mean of 110 ms with a standard deviation of 15 ms. What is the probability that the mean time (in milliseconds) for encoding 50 randomly selected images of this size will be between 104 and 115 ms?

4.1.11 Let $X_1, \ldots, X_n$ be independent discrete random variables identically distributed as:

$$f(x_i) = \begin{cases} 0.2, & x_i = 0, 1, 2, 3, 4, \\ 0, & \text{otherwise.} \end{cases}$$

Using the central limit theorem, find the approximate value of $P\left(\overline{X}_{100} > 2\right)$, where $\overline{X}_{100} = (1/100)\sum_{i=1}^{100} X_i$.

4.1.12. A population of disk drives manufactured by a certain company runs with a mean seek time of 10 ms with standard deviation of 1 ms. What proportion of samples of size 250 would you expect to result in a mean less than 9.9 ms?

4.1.13. Suppose that the national norm of a science test for 12th graders on a particular year has a mean of 215 and a standard deviation of 35.

(a) A random sample of 55 12th graders is selected. What is the probability that this group will average more than 225?

(b) A random sample of 200 12th graders is selected. What is the probability that this group will average over 225?

(c) A random sample of 35 12th graders is selected. What is the probability that this group will average over 225?

(d) How does the sample size influence the probability?

4.1.14. Scores on the Wechsler Adult Intelligence Scale for the 20 to 34 age group are approximately normally distributed with a mean of 110 and standard deviation of 25. If we select 100 people at random, what is the probability that this group will have an average score of 116 or above?

4.1.15. It is known that a healthy human body has an average temperature of 98.6°F, with a standard deviation of 0.95°F. Sixty healthy humans are selected at random. What is the probability that their temperature average is at least 98.8°F?

4.1.16. A random sample of size 100 is taken from a population with mean 1 and variance 0.04. Find the probability that the sample mean is between 0.99 and 1.

4.1.17. The lifetime X (in hours) of a certain electrical component has the probability density function (pdf) $f(x) = (1/3)e^{-(1/3)x}$, $x > 0$. If a random sample of 36 is taken from these components, find $P\left(\overline{X} < 2\right)$.

4.2 Sampling distributions associated with normal populations

The sampling distribution of a statistic will depend upon the population distribution from which the samples are taken. In this section we discuss the sampling distributions of some statistics that are based on a random sample drawn from a normal distribution. These statistics are used in many statistical procedures that are very important in solving real-world problems. The following result establishes the distribution of a linear combination of independent normal random variables.

Theorem 4.2.1 *Let* $X_1, \ldots, X_n$ *be independent random variables with the distribution of* X_i *being normal with mean* μ_i *and variance* σ_i^2. *Let* $a_1, a_2, \ldots, a_n$ *be real constants. Then the distribution of* $Y = \sum_{i=1}^n a_iX_i$ *is normal with mean* $\mu_Y = \sum_{i=1}^n a_i\mu_i$ *and variance* $\sigma_Y^2 = \sum_{i=1}^n a_i^2\sigma_i^2$.

Proof. The moment-generating (mgf) function of Y is given by

$$\begin{aligned}
M_Y(t) &= Ee^{\left(\sum_{i=1}^n a_iX_i\right)t} \\
&= \prod_i Ee^{(a_iX_i)t}, \quad \left[\text{by independence of } X_i's\right] \\
&= \prod_i Ee^{(a_it)X_i} \\
&= \prod_i M_{X_i}(a_it), \quad [\text{using the definition of mgf}] \\
&= \prod_i e^{\left(a_i\mu_it+(1/2)a_i^2\sigma_i^2t^2\right)}, \quad [\text{using mgf of a normal}] \\
&= e^{\left[\left(\sum_i a_i\mu_i\right)t+(1/2)\left(\Sigma_i a_i^2\sigma_i^2\right)t^2\right]},
\end{aligned}$$

which is the mgf of a normal random variable with mean $\sum_i a_i\mu_i$ and variance $\sum_i a_i^2\sigma_i^2$.

In Theorem 4.2.1 let $a_i = 1/n$, $\mu_i = \mu$, and $\sigma_1^2 = \sigma^2$, we obtain the following result, which provides the distribution of the sample mean.

Corollary 4.2.2 *Let* $X_1, \ldots, X_n$ *be a random sample of size* n *from a normal population with mean* μ *and variance* σ^2. *Then:*

$$\overline{X} = (1/n)\sum_{i=1}^n X_i$$

is normally distributed with mean $\mu_{\overline{X}} = \mu$ *and variance* $\sigma_{\overline{X}}^2 = \sigma^2/n$.

Recall that we have used the notation $X \sim N(\mu, \sigma^2)$ to mean that the random variable X is normally distributed with mean μ and variance σ^2. From Corollary 4.2.2, $\overline{X} \sim N(\mu, \sigma^2/n)$ and hence by the z-transformation we obtain the standard normal random variable, $Z = \left(\overline{X} - \mu\right)/(\sigma/\sqrt{n}) \sim N(0,1)$.

EXAMPLE 4.2.1

A company that manufactures cars claims that the gas mileage for its new line of hybrid cars, on the average, is 60 miles per gallon with a standard deviation of 4 miles per gallon. A random sample of 16 cars yielded a mean of 57 miles per gallon. If the company's claim is correct, what is the probability that the sample mean is less than or equal to 57 miles per gallon? Comment on the company's claim about the mean gas mileage per gallon of its cars. What assumptions did you make?

Solution

Let X *represent the gas mileage for the new car (in miles per gallon). If the company's claim is true, then from Corollary 4.2.2,* $\overline{X}$ *is normally distributed with mean* $\mu = 60$ *and variance* $\sigma^2/n = 16/16 = 1$*. Hence,*

$$P(\overline{X} \le 57) = P\left(\frac{\overline{X} - 60}{1} \le \frac{57 - 60}{1}\right)$$

$$= P(Z \le -3) \approx 1 - 0.999$$

$$= 0.001.$$

Therefore, if the company's claim is correct, it is very unlikely that the mean value of the random sample of 16 cars will be 57 miles per gallon. Because the mean is indeed 57 miles per gallon, we conclude that the company's claim is very likely not true. Here we have assumed that the sample of 16 measurements comes from a normal population, so that we could apply the results of Corollary 4.2.2.

Now we introduce some distributions that can be derived from a normal distribution. These distributions play a very important role in inferential statistics.

4.2.1 Chi-square distribution

A chi-square distribution is used in many inferential problems, for example, in inferential problems dealing with the variance. Recall that the chi-square distribution is a special case of a gamma distribution with $\alpha = n/2$ and $\beta = 2$. If n is a positive integer, then the parameter n is called the ***degrees of freedom***. However, if n is not an integer, but $\beta = 2$, we still refer to this distribution as a chi-square. The mgf of a χ^2- random variable is $M(t) = (1-2t)^{-n/2}$. The mean and variance of a chi-square distribution are $\mu = n$ and $\sigma^2 = 2n$, respectively. That is, the mean of a $\chi^2(n)$ random variable is equal to its degrees of freedom and the variance is twice the degrees of freedom. We now give some useful results for χ^2-random variables.

Theorem 4.2.3 *Let* $X_1, \ldots, X_k$ *be independent* χ^2- *random variables with* $n_1, \ldots, n_k$ *degrees of freedom, respectively. Then the sum* $V = \sum_{i=1}^{k} X_i$ *is chi-square distributed with* $n_1 + n_2 + \ldots + n_k$ *degrees of freedom.*

Proof. The mgf of V is

$$M_V(t) = \prod_{i=1}^{k} (1-2t)^{-n_i/2} = (1-2t)^{-\left(\sum_{i=1}^{k} n_i\right)/2}.$$

This implies that. $V \sim \chi^2\left(\sum_{i=1}^{k} n_i\right)$.

Our next result states that the difference of two chi-square random variables is a chi-square random variable, given by the following theorem. The proof is left as an exercise.

Theorem 4.2.4 *Let* X_1 *and* X_2 *be independent random variables. Suppose that* X_1 *is* χ^2 *with* n_1 *degrees of freedom, whereas* $Y = X_1 + X_2$ *is chi-square with* n *degrees of freedom, where* $n > n_1$*. Then* $X_2 = Y - X_1$ *is a chi-square random variable with* $n - n_1$ *degrees of freedom.*

The following result shows that we can generate a chi-square random variable from a gamma random variable.

Theorem 4.2.5 *If a random variable* X *has a gamma probability distribution with parameters* α *and* β*, then:*

$$Y = \frac{2X}{\beta} \sim \chi^2(2\alpha).$$

Proof. Recall that the mgf of the gamma random variable X is $(1-\beta t)^{-\alpha}$. Thus,

$$M_Y(t) = M_{\frac{2X}{\beta}}(t) = E\left(e^{\frac{2X}{\beta}t}\right)$$

$$= E\left(e^{X\left(\frac{2}{\beta}t\right)}\right) = M_X\left(\frac{2}{\beta}t\right)$$

$$= (1-2t)^{-\alpha} = (1-2t)^{-\frac{2\alpha}{2}}.$$

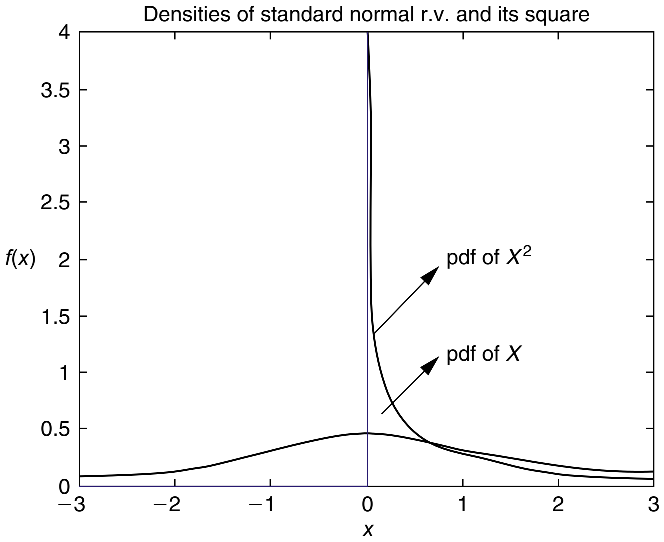


FIGURE 4.1 Probability density function (pdf) of standard normal random variable (r.v.) and the pdf of its square.

Hence, $Y \sim \chi^2(2\alpha)$.

The following result states that by squaring a standard normal random variable, we can generate a chi-square random variable, with 1 degree of freedom.

Theorem 4.2.6 *If* X *is a standard normal random variable, then* X^2 *is chi-square random variable with 1 degree of freedom.*

Proof. Because $X \sim N(0, 1)$, the mgf function of X^2 is

$$M_{X^2}(t) = \int_{-\infty}^{\infty} e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = (1-2t)^{-1/2}.$$

This implies that $X^2 \sim \chi^2(1)$. Fig. 4.1 gives the probability densities of the random variables X and X^2.

The following result is a direct consequence of Theorems 4.2.3 and 4.2.6. This result illustrates how to obtain a random sample from chi-square distribution if we have a random sample of n measurements from a normal population.

Theorem 4.2.7 *Let the random sample* $X_1, \ldots, X_n$ *be from an* $N(\mu,\sigma^2)$ *distribution. Then* $Z_i = (X_i - \mu)/\sigma$, $i = 1, \ldots, n$ *are independent standard normal random variables and*

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2,$$

has a χ^2 *distribution with* n *degrees of freedom. In particular, if* $X_1, \ldots, X_n$ *are independent standard normal random variables, then* $Y^2 = \sum_{i=1}^{n} X_i^2$ *is chi-square distributed with* n *degrees of freedom.*

If $X \sim \chi^2(n)$, then from the chi-square table, we can compute the values of $\chi^2_\alpha(n)$ such that:

$$P\left(X > \chi^2_\alpha(n)\right) = \alpha,$$

as shown by Fig. 4.2.

For example, if $X \sim \chi^2(15)$, to find $\chi^2_{0.95}(15)$ look in the chi-square table with the row labeled 15 degrees of freedom and the column headed $\chi^2_{0.950}$ and obtain the value as 7.26,094. Thus, with 15 degrees of freedom, $P(X > 7.26{,}094) =$ 0.95. Also, if X is a chi-square random variable with 11 degrees of freedom, from the chi-square table we have $\chi^2_{0.05}(11) =$ 19.675. Therefore, $P(X > 19.675) = 0.05$.

EXAMPLE 4.2.2

Let the random variables $X_1, X_2, \ldots, X_5$ be from an $N(5,1)$ distribution. Find a number a such that

$$P\left(\sum_{i=1}^{5} (X_i - 5)^2 \le a\right) = 0.90.$$

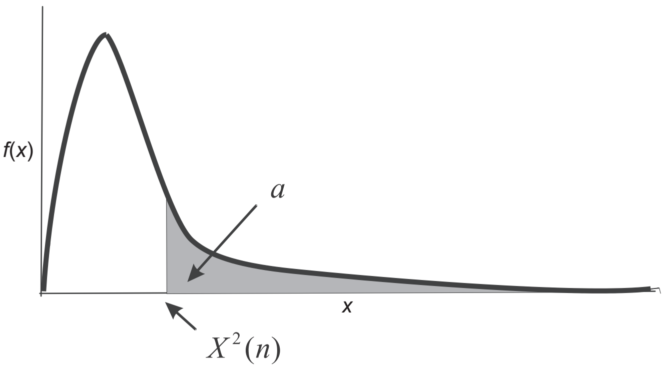


FIGURE 4.2 Chi-square probability density.

Solution

By Theorem 4.2.7, $\sum_{i=1}^{5} Z_i^2 = \sum_{i=1}^{5}\left(\frac{X_i-5}{1}\right)^2 = \sum_{i=1}^{5}(X_i-5)^2$ *has a chi-square distribution with 5 degrees of freedom.*

Because the upper tail area is 0.10, looking at the chi-square table with 5 degrees of freedom and the column corresponding to $\chi^2_{0.10}$, *we obtain a* = *9.23,635. Thus,*

$$P\left(\sum_{i=1}^{5}(X_i-5)^2 \le 9.23635\right) = 0.90.$$

EXAMPLE 4.2.3

Suppose that X is a χ^2—random variable with 20 degrees of freedom. Use the chi-square table to obtain the following:

(a) Find x_0 such that $P(X > x_0) = 0.95$.

(b) Find $P(X \le 12.443)$.

Solution

(a) *For 20 degrees of freedom, using the chi-square table, we have*:

$$P(X > 10.851) = 0.95.$$

Hence, $x_0 = 10.851$.

(b) *From the chi-square table,*

$$P(X \le 12.443) = 0.10.$$

The following result gives the probability distribution for a function of the sample variance S^2.

Theorem 4.2.8 *If* $X_1,\ldots,X_n$ *is a random sample from a normal population with the mean* μ *and variance* σ^2, *then*:

(a) *the random variable*

$\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$, *has a chi-square distribution with* (n − *1) degrees of freedom.*

(b) $\overline{X}$ *and* S^2 *are independent.*

Proof. We will prove only part (A). For (B), we will give some comments on the proof.

(a) We know from Theorem 4.2.7 that $(1/\sigma^2)\sum_{i=1}^{n}(X_i-\mu)^2$ has a chi-square distribution with n degrees of freedom. Thus,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\overline{X}+\overline{X}-\mu\right)^2$$

$$= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{n}\left(\overline{X}-\mu\right)^2\right]$$

$$\left(\text{Since}\quad 2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(\overline{X}-\mu\right) = 0\right)$$

$$= \frac{(n-1)S^2}{\sigma^2}+\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2.$$

The left-hand side of this equation has a chi-square distribution with n degrees of freedom. Also, since $\left(\overline{X}-\mu\right)/(\sigma|\sqrt{n}\,)\sim N(0,1)$ by Theorem 4.2.6 we have $\left[\left(\overline{X}-\mu\right)\big|(\sigma|\sqrt{n}\,)\right]^2\sim\chi^2(1)$. Now from Theorem 4.2.4, $(n-1)\ S^2/\sigma^2 \sim \chi^2\ (n-1)$.

(b) We will accept the result of (B) without proof here. A rigorous proof depends on the geometric properties of the multivariate normal distribution, which is beyond the scope of this book. A proof based on mgf functions is relatively straightforward, where essentially we can first show that the random variable $\overline{X}$ and the vector of random variables $\left(X_1-\overline{X},\ldots,X_n-\overline{X}\right)$ are independent. Because S^2 is a function of the vector $\left(X_1-\overline{X},\ldots,X_n-\overline{X}\right)$, it is then independent of $\overline{X}$.

EXAMPLE 4.2.4

Let X_1, X_2, …, X_{10} be a random sample from a normal distribution with $\sigma^2 = 0.8$. Find two positive numbers a and b such that the sample variance S^2 satisfies

$$P\left(a\le S^2\le b\right) = 0.90.$$

Solution

Because $\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$, *we have*

$$P\left(a\le S^2\le b\right) = P\left(\frac{(n-1)a}{\sigma^2}\le\frac{(n-1)S^2}{\sigma^2}\le\frac{(n-1)b}{\sigma^2}\right).$$

The desired values can be found by setting the upper tail area and lower tail area each equal to 0.05. Using the chi-square table with n − *1 = 9 degrees of freedom, we have:*

$$\frac{(n-1)b}{\sigma^2} = \frac{9b}{0.8} = 16.919 = \chi^2_{0.05,9},$$

which implies b *= ((16.919) × (0.8)/9) = 1.50. Similarly,*

$$\frac{(n-1)a}{\sigma^2} = \frac{9a}{0.8} = 3.325 = \chi^2_{0.95,9}.$$

So we have a *= ((3.325) × (0.8)/9) = 0.295.*
Hence,

$$P\left(0.295\le S^2\le 1.50\right) = 0.90.$$

It is important to note that this is not the only interval that would satisfy:

$$P\left(a\le S^2\le b\right) = 0.90,$$

but it is a convenient one.

EXAMPLE 4.2.5

A fruit-drink company wants to know the variation, as measured by the standard deviation, of the amount of juice in 16-oz cans. From past experience, it is known that $\sigma^2 = 2$. The company statistician decides to take a sample of 25 cans from the production line and compute the sample variance. Assuming that the sample values may be viewed as a random sample from a normal population, find a value of b such that $P(S^2 > b) = 0.05$.

Solution

To find the necessary probability, use the fact that $(n - 1)S^2/\sigma^2 \sim \chi^2(n-1)$, *with* $n = 25$,

$$0.05 = P(S^2 > b) = P\left(\frac{24S^2}{2} > \frac{24b}{2}\right)$$

$$= P(\chi^2 > c).$$

From the chi-square table we obtain, $c = 36.4151$. *Hence,* $b = \frac{2}{24}c = \frac{2}{24}(36.4151) = 3.03$ *and*

$$P(S^2 > 3.03) = 0.05.$$

Summary of Chi-Square Distribution

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ random variables. Then

1. $\overline{X}$ has $N(\mu, \sigma^2/n)$ distribution,
2. $(n-1)S^2/\sigma^2$ has a chi-square distribution with $(n-1)$ degrees of freedom, and
3. $\overline{X}$ and S^2 are independent.
4. A χ^2- random variable has a mean equal to its degrees of freedom and a variance equal to twice its degrees of freedom.

4.2.2 Student *t* distribution

Let the random variables $X_1, \ldots, X_n$ follow a normal distribution with mean μ and variance σ^2. If σ is known, then we know that $\sqrt{n}\left(\left(\overline{X}-\mu\right)/\sigma\right)$is $N(0,1)$. However, if σ is not known (as is usually the case), then it is routinely replaced by the sample standard deviation s. If the sample size is large, one could suppose that $s \approx \sigma$ and apply the central limit theorem and obtain that $\sqrt{n}\left(\left(\overline{X}-\mu\right)/S\right)$ is approximately an $N(0,1)$. However, if the random sample is small, then the distribution of $\sqrt{n}\left(\left(\overline{X}-\mu\right)/S\right)$ is given by the so-called *Student distribution* (or simply *t* distribution). This was originally developed by W. S. Gosset in 1908. Because his employer, the Guinness brewery, would not permit him to publish this important work in his own name, he used the pseudonym "Student." Thus, the distribution is known as the Student *t* distribution.

Definition 4.2.2 *If* Y *and* Z *are independent random variables,* Y *has a chi-square distribution with* n *degrees of freedom, and* Z $\sim$ N*(0, 1), then*:

$$T = \frac{Z}{\sqrt{Y/n}}$$

is said to have a (Student) **t-distribution** *with* n *degrees of freedom. We denote this by* $T \sim T_n$.

The probability density of the random variable T with n degrees of freedom is given by:

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, -\infty < t < \infty.$$

Fig. 4.3 illustrates the behavior of the *t* distributions for $n = 2$, 10, 20, and 30. It is clear from Fig. 4.3 that as n becomes larger and larger, it is almost impossible to distinguish the graphs. It can be shown that the *t* distribution tends to a standard normal distribution as the degrees of freedom (equivalently, the sample size n) tend to infinity. In fact, *the standard normal*

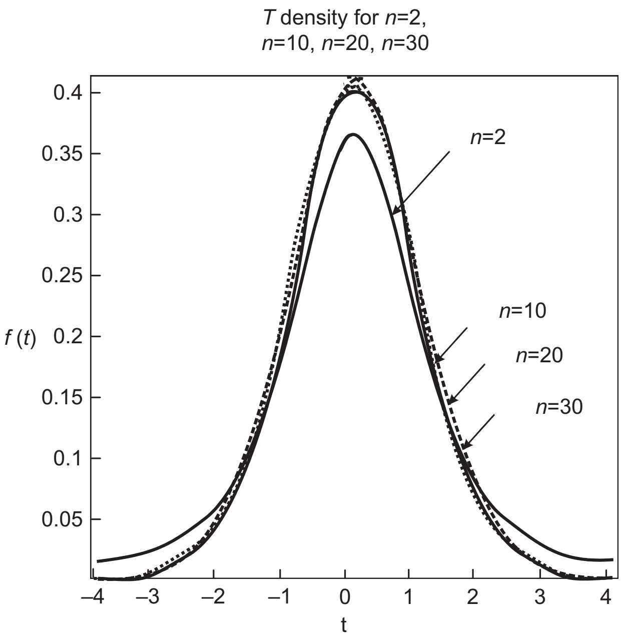


FIGURE 4.3 The Student *t* distribution.

distribution provides a good approximation to the t-*distribution for sample sizes of 30 or more*. We will use this approximation in the statistical inference problems for $n \geq 30$.

The t density is symmetric about zero, and then we have $E(T) = 0$. If $n > 2$, it can be shown that $Var\ (T) = n/(n-2)$. The value of $t_{\alpha,n}$ is such that $P\ (t > t_{\alpha,n}) = \alpha$ (the shaded area in Fig. 4.4) is obtained from the t table. For example, if a random variable X has a t distribution with 9 degrees of freedom and $\alpha = 0.01$, then $t_{0.01,9} = 2.821$.

If we have a random sample from a normal population, the following result involving a t distribution is useful in applications.

Theorem 4.2.9 *If* $\overline{\mathrm{X}}$ *and* S^2 *are the mean and the variance of a random sample of size* n *from a normal population with mean* μ *and variance* σ^2*, then:*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

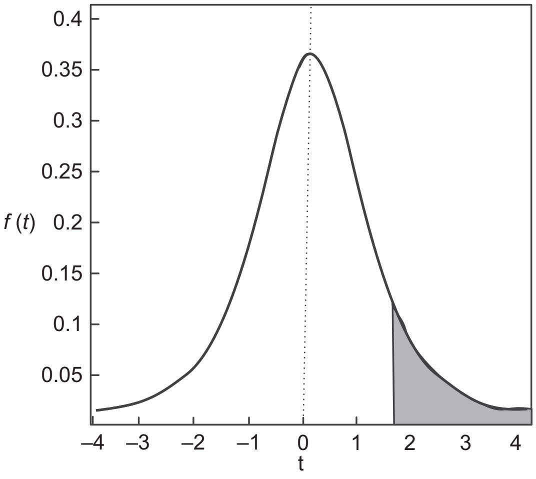


FIGURE 4.4 Probability of *t* distribution.

has a t *distribution with* (n − 1) *degrees of freedom.*

Proof. By Corollary 4.2.2,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

By Theorem 4.2.8, we have:

$$Y = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sim \chi^2(n-1).$$

Hence,

$$T = \frac{\dfrac{\overline{X} - \mu}{(\sigma/\sqrt{n})}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}} \sim \frac{Z}{\sqrt{\dfrac{\chi^2(n-1)}{n-1}}}.$$

Also, $\overline{X}$ and S^2 are independent. Thus, Y and Z are independent, and by Definition 4.2.2, T follows a t distribution with $(n-1)$ degrees of freedom.

How can we distinguish between given degrees of freedom and the degrees of freedom from a sample? For the t distribution, if n is given as the degrees of freedom, we will just use n. However, if a random sample of size n is given, then the corresponding degrees of freedom will be $(n-1)$, as given in Theorem 4.2.9.

The assumption that the sample comes from a normal population is not that onerous. In practice, it is necessary to check that the sampled population is approximately bell shaped and not too skewed. Construction of the normal-scores plot or histogram is a way to check for approximate normality. See Project 4C.

EXAMPLE 4.2.6

A manufacturer of fuses claims that with 20% overload, the fuses will blow in less than 10 minutes on average. To test this claim, a random sample of 20 of these fuses was subjected to a 20% overload, and the times it took them to blow had a mean of 10.4 minutes and a sample standard deviation of 1.6 minutes. It can be assumed that the data constitute a random sample from a normal population. Do they tend to support or refute the manufacturer's claim?

Solution

Given $\overline{y} = 10.4$, $s = 1.6$, $n = 20$, *and* $\mu = 10$. *Hence,*

$$t = \frac{\overline{y} - \mu}{s/\sqrt{n}} = \frac{10.4 - 10}{1.6/\sqrt{20}} = 1.118.$$

The degree of freedom is $n - 1 = 19$. *From the* t*-table, the probability that* t *exceeds 1.328 is 0.10, and because the observed value of* $t = 1.118$ *is less than* $t_{0.10,19} = 1.328$ *and* 0.10 *is a pretty large probability, we conclude that the data tend to agree with the manufacturer's claim.*

We will study the problems of the foregoing type in Chapter 6, where we will be learning about hypothesis testing. Prior to Gosset's work on the t distribution, a very large number of observations were necessary for the design and analysis of experiments. Today, the use of the t distribution often makes it possible to draw reliable conclusions from samples as small as 15 to 30 experimental units, provided that the samples are representative of their populations and that normality could reasonably be assumed or justified for the population. Example 4.2.7 suggests that we need to be careful about the use of t distribution. It depends not only on sample size, but also on the knowledge deviation.

EXAMPLE 4.2.7

The human gestation period—the period of time between conception and labor—is approximately 40 weeks (280 days), measured from the first day of the mother's last menstrual period. For a newborn full-term infant, the appropriate length for gestational age is assumed to be normally distributed with $\mu = 50$ cm and $\sigma = 1.25$ cm. Compute the probability that a random sample of 20 infants born at full term results in a sample mean greater than 52.5 cm.

Solution

Let X *be the length (measured in centimeters) of a newborn full-term infant. Then* $\overline{X} \sim N(50, 1.56\,/20)$. *Note that even though the sample size is small, since* σ *is known, we do not use* t *distribution, instead we use normal distribution. Hence,*

$$P(\overline{X} > 52.5) = P\left(z > \frac{52.5 - 50}{1.25/\sqrt{20}} = 8.94\right) \approx 0.$$

Thus, the probability of such an occurrence is negligible.

In the previous example, it should be noted that $P(\overline{X} > 52.5) \approx 0$ does not imply that the probability of observing a newborn full-term infant with length greater than 52.5 cm is zero. In fact, with 19 degrees of freedom, $P(X > 52.5) = P(Z > 2) \approx 0.0228$.

4.2.3 *F*-distribution

The *F*-distribution was developed by Fisher to study the behavior of two variances from random samples taken from two independent normal populations. In applied problems we may be interested in knowing whether the population variances are equal, based on the response of the random samples. Knowing the answer to such a question is also important in selecting the appropriate statistical methods to study their true means.

Definition 4.2.3 *Let* U *and* V *be chi-square random variables with* n_1 *and* n_2 *degrees of freedom, respectively. Then if* U *and* V *are independent,*

$$F = \frac{U/n_1}{V/n_2},$$

is said to have an **F-distribution** *with* n_1 *numerator degrees of freedom and* n_2 *denominator degrees of freedom. We denote this by* $F \sim F\,(n_1, n_2)$.

The pdf for a random variable $X \sim F\,(n_1, n_2)$ is given by:

$$f(x) = \begin{cases} \dfrac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)}\left(\dfrac{n_1}{n_2}\right)^{n_1/2} x^{\frac{n_1}{2}-1}\left(1+\dfrac{n_1}{n_2}x\right)^{-(n_1+n_2)/2}, & x > 0 \\ 0, & \text{elsewhere.} \end{cases}$$

A graph of *f*(*x*) for various values of *n* is given in Fig. 4.5.

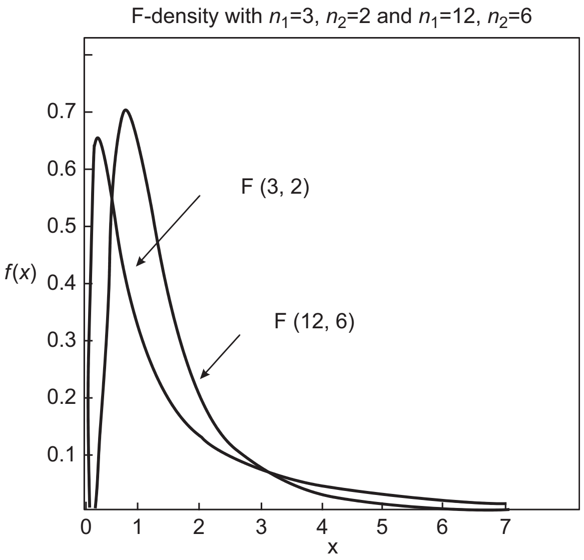


FIGURE 4.5 Probability density functions of the *F*-distribution.

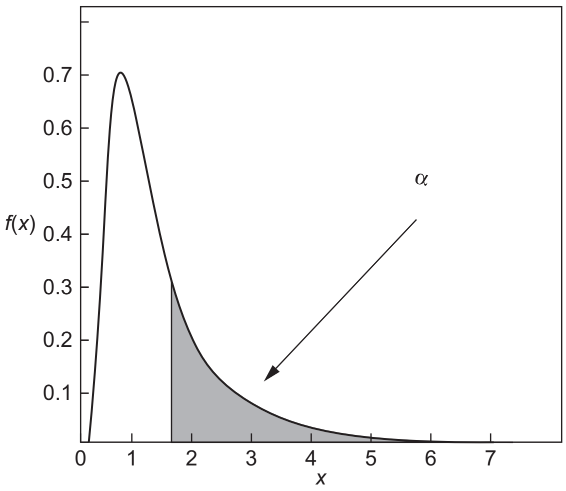


FIGURE 4.6 Probability density functions of F-distribution.

To find $F_a\,(n_1,n_2)$ such that $P(F > F_a\,(n_1,n_2)) = \alpha$ (shaded area in Fig. 4.6), we use the F table. For example, if F has 3 numerator and 6 denominator degrees of freedom, then $F_{0.01}\,(3, 6) = 9.78$.

If we know $F_a\,(n_1, n_2)$, it is possible to find $F_{1-\alpha}\,(n_2, n_1)$ by using the identity

$$F_{1-\alpha}(n_2,\, n_1) = 1/F_\alpha(n_1,\, n_2).$$

Using this identity, we can obtain $F_{0.99}\,(6, 3) = 1/F_{0\cdot 01}\,(3, 6) = 1/9.78 = 0.10225$.

When we need to compare the variances of two normal populations, we will use the following result.

Theorem 4.2.10 *Let two independent random samples of size* n_1 and n_2 *be drawn from two normal populations with variances* σ_1^2, σ_2^2, *respectively. If the variances of the random samples are given by* $\mathrm{S}_1^2, \mathrm{S}_2^2$, *respectively, then the statistic*:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2},$$

has the F *distribution with* $(\mathrm{n}_1 - 1)$ *numerator and* $(\mathrm{n}_2 - 1)$ *denominator degrees of freedom.*

Proof. From Theorem 4.2.9, we know that:

$$U = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$$

and

$$V = \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1).$$

Also, U and V are independent. From Definition 4.2.3, $F \sim F\,(n_1 - 1,\, n_2 - 1)$.

Corollary 4.2.11 If $\sigma_1^2 = \sigma_2^2$, then

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1).$$

when $\sigma_1^2 = \sigma_2^2$, we refer to them as two populations that are homogeneous with respect to their variances.

EXAMPLE 4.2.8

Let S_1^2 denote the sample variance for a random sample of size 10 from population I and let S_2^2 denote the sample variance for a random sample of size 8 from population II. The variance of population I is assumed to be three times the variance of population II. Find two numbers a and b such that $P\left(a \le S_1^2/S_2^2 \le b\right) = 0.90$ assuming S_1^2 to be independent of S_2^2.

Solution

From the problem, we can assume that $\sigma_1^2 = 3\sigma_2^2$ *with* $n_1 = 10$ *and* $n_2 = 8$*. Thus, we can write*:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2/3\sigma_2^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{3S_2^2},$$

which has F*-distribution with* $n_1 - 1 = 9$ *numerator and* $n_2 - 1 = 7$ *denominator degrees of freedom. Using the F-table,* $F_{0.05}$ *(9, 7)* = *3.68. Now to find* $F_{0.95}$ *such that*:

$$P\left(\frac{S_1^2}{3S_2^2} < F_{0.95}\right) = 0.05.$$

We proceed as follows:

$$P\left(\frac{S_1^2}{3S_2^2} < F_{0.95}\right) = P\left(\frac{3S_2^2}{S_1^2} > \frac{1}{F_{0.95}}\right) = 0.05.$$

Indexing $v_1 = 7$ *and* $v_2 = 9$ *in the F-table, we have* $1/F_{0.95}(7,9) = F_{0.05}(9,7) = 3.68$ *or* $F_{0.95} = 1/3.68 = 0.2717$*. Hence, the entire probability statement is given by*:

$$P\left(0.2717 \le \frac{S_1^2}{3S_2^2} \le 3.68\right) = P\left(0.815 \le \frac{S_1^2}{S_2^2} \le 11.04\right) = 0.90.$$

Thus, $a = 0.815$ *and* $b = 11.04$*.*

When we need to find values not given in a t-table (or chi-square or F-tables), and those values are not available in the corresponding table, what can we do? Most of the time, it is easier to use statistical programs to obtain so-called P values (introduced in Chapter 6). In similar situations, most of the values given in a solution manual are obtained in this way. If the software is not available and we need to find these values from one of these (t-, chi-square, or F-) tables, then, using either linear interpolation or transformations, we can obtain approximate values. We will illustrate this only for linear interpolation. Given two points (x_1, y_1) and (x_2, y_2) on a line, any other point (x, y) on the line satisfies the following equation

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1).$$

Using this relationship, let us say, we want to find the t-value for $\alpha = 0.15$ with 6 degrees of freedom. In our table we have t values for $\alpha = 0.1$ and $\alpha = 0.25$. Then,

$$t_{0.15,6} \approx 1.439756 + \frac{(0.717558 - 1.439756)}{(0.25 - 0.1)}(0.15 - 0.1)$$

$$= 1.199023.$$

We will use this method for finding critical values of t, chi-square, and F-values that are not available in the table.

Exercises 4.2

4.2.1. Let Y have a chi-square distribution with 15 degrees of freedom. Find the following probabilities.
(a) $P(Y \le y_0) = 0.025$.
(b) $P\ (a < Y < b) = 0.95$.
(c) $P(Y \ge 22.307)$.

4.2.2. Let Y have a chi-square distribution with 7 degrees of freedom. Find the following probabilities.
(a) $P(Y > y_0) = 0.025$
(b) $P\ (a < Y < b) = 0.90$
(c) $P(Y > 1.239)$.

4.2.3. The time to failure T of a microwave oven has an exponential distribution with pdf:

$$f(t) = \frac{1}{2}e^{-t/2}, \quad t > 0.$$

If three such microwave ovens are chosen and $\bar{t}$ is the mean of their failure times, find the following:

(a) Distribution of $\overline{T}$.

(b) $P(\overline{T} > 2)$.

4.2.4. Let $X_1, X_2, \ldots, X_{10}$ be a random sample from a standard normal distribution. Find the numbers a and b such that:

$$P\left(a \leq \sum_{i=1}^{10} X_i^2 \leq b\right) = 0.95.$$

4.2.5. Let $X_1, X_2, \ldots, X_5$ be a random sample from the normal distribution with mean 55 and variance 223. Let

$$Y = \sum_{i=1}^{5} (X_i - 55)^2/223$$

and

$$Z = \sum_{i=1}^{5} \left(X_i - \overline{X}\right)^2/223.$$

(a) Find the distribution of the random variables Y and Z.

(b) Are Y and Z independent?

(c) Find (i) $P(0.554 \leq Y \leq 0.831)$, and (ii) $P(0.297 \leq Z \leq 0.484)$.

4.2.6. Let X and Y be independent chi-square random variables with 14 and 5 degrees of freedom, respectively. Find:

(a) $P(|X-Y| \leq 11.15)$,

(b) $P(|X-Y| \geq 3.8)$.

4.2.7. A particular type of vacuum-packed coffee packet contains an average of 16 oz. It has been observed that the number of ounces of coffee in these packets is normally distributed with $\sigma = 1.41$ oz. A random sample of 15 of these coffee packets is selected, and the observations are used to calculate s. Find the numbers a and b such that $P(a \leq S^2 \leq b) = 0.90$.

4.2.8. An optical firm buys glass slabs to be ground into lenses, and it is known that the variance of the refractive index of the glass slabs is to be no more than 1.04×10^{-3}. The firm rejects a shipment of glass slabs if the sample variance of 16 pieces selected at random exceeds 1.15×10^{-3}. Assuming that the sample values may be looked on as a random sample from a normal population, what is the probability that a shipment will be rejected even though $\sigma^2 = 1.04 \times 10^{-3}$?

4.2.9. Assume that T has a t distribution with 8 degrees of freedom. Find the following probabilities.

(a) $P(T \leq 2.896)$.

(b) $P(T \leq -1.860)$.

(c) The value of a such that $P(-a < T < a) = 0.99$.

4.2.10. Assume that T has a t distribution with 15 degrees of freedom. Find the following probabilities.

(a) $P(T \leq 1.341)$.

(b) $P(T \geq -2.131)$.

(c) The value of a such that $P(-a < T < a) = 0.95$.

4.2.11. A psychologist claims that the mean age at which female children start walking is 11.4 months. If 20 randomly selected female children are found to have started walking at a mean age of 12 months with standard deviation of 2 months, would you agree with the psychologist's claim? Assume that the sample came from a normal population.

4.2.12. Let U_1 and U_2 be independent random variables. Suppose that U_1 is χ^2 with v_1 degrees of freedom while $U = U_1 + U_2$ is chi-square with v degrees of freedom, where $v > \mathrm{v}_1$. Then prove that U_2 is a chi-square random variable with $v - v_1$ degrees of freedom.

4.2.13. Show that if $X \sim \chi^2\,(v)$, then $EX = v$ and $Var\,(X) = 2v$.

4.2.14. Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim \chi^2\,(1)$, for $i = 1, \ldots, n$. Show that the distribution of

$$Z = \frac{\overline{X} - 1}{\sqrt{2/n}}$$

as $n \to \infty$ is standard normal.

4.2.15. Find the variance of S^2, assuming the sample $X_1, X_2, \ldots, X_n$ is from $N(\mu, \sigma^2)$.

4.2.16. Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with parameter θ. Show that the random variable $2\theta^{-1}\left(\sum_{i=1}^{n} X_i\right) \sim \chi^2(2n)$.

4.2.17. Let X and Y be independent random variables from an exponential distribution with common parameter $\theta = 1$. Show that X/Y has an F distribution. What is the number of the degrees of freedom?

4.2.18. Prove that if X has a t distribution with n degrees of freedom, then $X^2 \sim \mathrm{F}\ (1, n)$.

4.2.19. Let X be F distributed with 9 numerator and 12 denominator degrees of freedom. Find

(a) $P(X \le 3.87)$.

(b) $P(X \le 0.196)$.

(c) The value of a and b such that $P\ (a < Y < b) = 0.95$.

4.2.20. Prove that if $X \sim F(n_1,n_2)$, then $1/X \sim F(n_2,n_1)$.

4.2.21. Find the mean and variance of $F(n_1, n_2)$ random variable.

4.2.22. Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample with sample mean $\overline{X}_1$ from a normal population with mean μ_1 and variance σ_1^2, and let $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample with sample mean $\overline{X}_2$ from a normal population with mean μ_2 and variance σ_2^2. Assume the two samples are independent. Show that the sampling distribution of $\left(\overline{X}_1 - \overline{X}_2\right)$ is normal with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$.

4.2.23. Let $X_1, X_2, \ldots, X_{n1}$ be a random sample from a normal population with mean μ_1 and variance σ^2, and $Y_1, Y_2, \ldots, Y_{n2}$ be a random sample from an independent normal population with mean μ_2 and variance σ^2. Show that

$$T = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim T_{(n_1+n_2-2)}.$$

4.2.24. Show that a t distribution tends to a standard normal distribution as the degrees of freedom tend to infinity.

4.2.25. Show that the mgf of a χ^2 random variable with n degrees of freedom is $M(t)=(1 - 2t)^{-n/2}$. Using the mgf, show that the mean and variance of a chi-square distribution are n and $2n$, respectively.

4.2.26. Let the random variables $X_1, X_2, \ldots, X_{10}$ be normally distributed with mean 8 and variance 4. Find a number a such that

$$P\left(\sum_{i=1}^{10}\left(\frac{X_i - 8}{2}\right)^2 \le a\right) = 0.95.$$

4.2.27. Let $X^2 \sim F(1,n)$. Show that $X \sim t(n)$.

4.3 Order statistics

In practice, the random variables of interest may depend on the relative magnitudes of the observed variable. For example, we may be interested in the maximum mileage per gallon of a particular class of cars. In this section, we study the behavior of ordering a random sample from a continuous distribution.

Definition 4.3.1 *Let* $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}$ *be a random sample from a continuous distribution with pdf* f(x). *Let* $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ *be a permutation of* $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}$ *such that*

$$Y_1 \le Y_2 \le \cdots \le Y_n.$$

Then the ordered random variables $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ *are called the* **order statistics** *of the random sample* $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}$. *Here* Y_k *is called the* **kth order statistic.** *Because of continuity, the equality sign could be ignored.*

Remark. Although $X_i's$ are iid random variables, the random variables $\mathrm{Y}_i's$ are neither independent nor identically distributed.

Thus, the minimum of $\mathrm{X}_i's$ is

$$Y_1 = \min\ (X_1, \ldots, X_n)$$

and the maximum is

$$Y_n = \max\ (X_1, \ldots, X_n).$$

The order statistics of the sample X_1, X_2, …, X_n can also be denoted by $X_{(1)}$, $X_{(2)}$, …, $X_{(\mathrm{n})}$ where

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}.$$

Here, $X_{(k)}$ is the kth order statistic and is equal to Y_k in Definition 4.3.1. One of the most commonly used order statistics is the median, the value in the middle position in the sorted order of the values.

EXAMPLE 4.3.1

(i) The range $R = Y_n - Y_1$ is a function of order statistics.
(ii) The sample median M equals Y_{m+1} if $n = 2m + 1$.

Hence, the sample median M is an order statistic, when n is odd. If n is even, then the sample median can be obtained using the order statistic $M = (1/2)\ [Y_{n/2} + Y_{(n/2)+1}]$.

The following result is useful in determining the distribution of functions of more than one order statistics.

Theorem 4.3.1 *Let* X_1, …, X_n *be a random sample from a population with pdf* $\mathrm{f(x)}$. *Then the joint pdf of order statistics* $\mathrm{Y}_1,\ldots,\mathrm{Y}_\mathrm{n}$ *is*:

$$f(y_1,\ldots,y_n) = \begin{cases} n!f(y_1)f(y_2)\ldots f(y_n), & \text{for } y_1 < \cdots < y_n \\ 0, & \text{otherwise.} \end{cases}$$

The pdf of the kth order statistic is given by the following theorem.

Theorem 4.3.2 *The pdf of* $\mathrm{Y_k}$ *is*:

$$f_k(y) = f_{Y_k}(y) = \frac{n!}{(k-1)!(n-k)!} f(y)\ (F(y))^{k-1}(1-F(y))^{n-k},$$

for $-\infty < \mathrm{y} < \infty$, *where* $\mathrm{F(y) = P(X_i \leq y)}$ *is the cumulative distribution function (cdf) of* $\mathrm{X_i}$.

In particular, the pdf of Y_1 is $f_1\ (y) = nf\ (y)\ [1{-}F(y)]^{n-1}$ and the pdf of Y_n is $\mathrm{f}_n(y) = nf(y)[F(y)]^{n-1}$. In the following example, we will derive the pdf for Y_n.

EXAMPLE 4.3.2

Let X_1, …, X_n be a random sample from $U[0,1]$. Find the pdf of the kth order statistic Y_k.

Solution

Since the pdf of $\mathrm{X_i}$ *is* $\mathrm{f(x) = 1, 0 \leq x \leq 1}$, *the cdf is* $\mathrm{F(x) = x,\ 0 \leq x \leq 1}$. Using Theorem 4.3.2, *the pdf of the* k*th order statistic* Y_k *reduces to*:

$$f_k(y) = \frac{n!}{(k-1)!(n-k)!} y^{k-1}(1-y)^{n-k},\ \ 0 \leq y \leq 1$$

which is a beta distribution with $\alpha = \mathrm{k}$ *and* $\beta = \mathrm{n} - \mathrm{k} + 1$.

The next example gives the so-called extreme (i.e., largest) value distribution, which is the distribution of the order statistic Y_n.

EXAMPLE 4.3.3

Find the distribution of the nth order statistic Y_n of the sample X_1, …, X_n from a population with pdf $f(x)$.

Solution

Let the cdf of $\mathrm{Y_n}$ *be denoted by* $\mathrm{F_n(y)}$. *Then:*

$$\begin{aligned} F_n(y) &= P(Y_n \leq y) = P\left(\max_{1\leq i\leq n} X_i \leq y\right) \\ &= P(X_1 \leq y,\ldots,X_n \leq y) = [F(y)]^n(\text{by independence}). \end{aligned}$$

Hence, the pdf $f_n(y)$ *of* Y_n *is:*

$$f_n(y) = \frac{d}{dy}[F(y)]^n = n[F(y)]^{n-1}\frac{d}{dy}F(y)$$

$$= n[F(y)]^{n-1}f(y).$$

In particular, if $X_1, \ldots, X_n$ *is a random sample from U[0, 1], then the cumulative extreme value distribution is given by*:

$$F_n(y) = \begin{cases} 0, & y < 0 \\ y^n, & 0 \le y \le 1 \\ 1, & y > 1. \end{cases}$$

EXAMPLE 4.3.4

A string of 10 light bulbs is connected in series, which means that the entire string will not light up if any one of the light bulbs fails. Assume that the lifetimes of the bulbs, $\tau_1, \ldots, \tau_{10}$, are independent random variables that are exponentially distributed with mean 2. Find the distribution of the life length of this string of light bulbs.

Solution

Note that the pdf of τ_i *is* $f(t) = 2e^{-2t}$, $0 < t < \infty$, *and the cumulative distribution of* τ_i *is* $F_{\tau i}(t) = 1-e^{-2t}$. *Let* T *represent the lifetime of this string of light bulbs. Then,*

$$T = \min(\tau_1, \ldots, \tau_{10}).$$

Thus,

$$F_T(t) = 1-[1-F_{\tau_i}(t)]^{10}.$$

Hence, the density of T *is obtained by differentiating* $F_T(t)$ *with respect to* t, *that is,*

$$f_T(t) = 10f_{\tau_i}(t)[1 - F_{\tau_i}(t)]^9$$

$$= \begin{cases} 2(10)e^{-2t}\left(e^{-2t}\right)^9 = 20e^{-20t}, & 0 < t < \infty \\ 0, & \text{otherwise.} \end{cases}$$

The joint pdf of the order statistics is given by the following result.

Theorem 4.3.3 *Let* $X_1, \ldots, X_n$ *be a random sample with continuous pdf* f(x) *and a distribution function* F(x). *Let* $Y_1, \ldots, Y_n$ *be the order statistics. Then for any* $1 \le i < k \le n$ *and* $-\infty < x \le y < \infty$, *the joint pdf of* Y_i *and* Y_k *is given by*:

$$f_{Y_i,Y_k}(x, y) = \frac{n!}{(i-1)!(k-i-1)!(n-k)!}[F(x)]^{i-1}$$

$$\times [F(y) - F(x)]^{k-i-1}[1 - F(y)]^{n-k}f(x)f(y)$$

EXAMPLE 4.3.5

Let $X_1, \ldots, X_n$ be a random sample from U[0,1]. Find the joint pdf of Y_2 and Y_5.

Solution

Taking i = 2 *and* k = 5 *in* Theorem 4.3.3, *we get the joint pdf of* Y_2 *and* Y_5 *as*:

$$f_{Y_2,Y_5}(x,y) = \frac{n!}{(2-1)!(5-2-1)!(n-5)!}[F(x)]^{2-1}$$

$$[F(y)-F(x)]^{5-2-1} \times [1-F(y)]^{n-5} f(x)f(y)$$

$$= \begin{cases} \dfrac{n!}{2(n-5)!}x(y-x)^2(1-y)^{n-5}, \ 0 < x \leq y < 1 \\ 0, \qquad\qquad\qquad \text{otherwise.} \end{cases}$$

Exercises 4.3

4.3.1. The lifetime X of a certain electrical fuse has the following pdf:

$$f(x) = \begin{cases} \dfrac{1}{10}e^{-x/10}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Suppose two such fuses are in series and operate independently in a system. Find the pdf of the lifetime Y of the system. (The system will work only if both of the fuses operate.)

4.3.2. Suppose that time between two telephone calls at an office, in minutes, is uniformly distributed on the interval [0, 20]. If there were 15 calls, **(i)** what is the probability that the longest time interval between calls is less than 15 minutes? **(ii)** What is the probability that the shortest time interval between calls is greater than 5 minutes?

4.3.3. Let X_1, X_2, X_3 be three random variables of discrete type. Let X_1, X_2 take values 0, 1, and X_3 take values 1, 2, 3. What are the values of Y_1, Y_2, Y_3?

4.3.4. Let X_1, …, X_{10} be a random sample from $U[0, 1]$. Find the joint density of Y_2 and Y_7, where Y_i, $\mathrm{i} = 1, 2, \ldots, 10$ are order statistics of X_1, …, X_{10}.

4.3.5. Let X_1, …,X_n be a random sample from exponential distribution with a mean of θ. Show that $Y_1 = \min(X_1, X_2, \ldots, X_n)$ has an exponential distribution with mean θ/n. Also, find the pdf of $\mathrm{Y}_n = \max(X_1, X_2, \ldots,X_n)$.

4.3.6. A string of 10 light bulbs is connected in parallel, which means that the entire string will fail to light up only if all 10 of the light bulbs fail. Assume that the lifetimes of the bulbs, τ_1, …, τ_{10}, are independent random variables that are exponentially distributed with mean θ. Find the distribution of the lifetimes of this string of light bulbs.

4.3.7. Let X_1, …, X_n be a random sample from the uniform distribution $f(x) = 1/2$, $0 \leq x \leq 2$. Find the pdf for the range $R = (X_{(n)} - X_{(1)})$.

4.3.8. Given a sample of 25 observations from a distribution with pdf:

$$f(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

let M be the sample median. Compute $P(M \geq (b)$.

[Hint: Note that M is the 13th order statistic.]

4.3.9. Let X_1, …, X_n be a random sample from a normal population with mean 10 and variance 4. What is the probability that the largest observation is greater than 10?

4.3.10. Let X_1, …, X_n be a random sample from an exponential population with parameter θ. Let Y_1, …, Y_n be the ordered random variables.

(a) Show that the sampling distributions of Y_1 and Y_n are given by

$$f_1(y_1) = \begin{cases} \dfrac{n}{\theta}e^{-ny_1/\theta}, \text{ if } & y_1 > 0 \\ 0, & \text{otherwise,} \end{cases}$$

and

$$f_n(y_n) = \begin{cases} \frac{n}{\theta}e^{-y_n/\theta}\left[1 - e^{-y_n/\theta}\right]^{n-1}, \text{ if } & y_n > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(b) Let $n = 2l + 1$. Show that the sampling distribution of the median, M, is given by:

$$f(m) = \begin{cases} \frac{n!}{(l!)^2\theta}e^{-m(l+1)/\theta}\left[1 - e^{-m/\theta}\right]^l, & \text{for } m > 0 \\ 0, & \text{otherwise.} \end{cases}$$

4.3.11. Let $X_1, \ldots, X_n$ be a random sample from a beta distribution with $\alpha = 2$ and $\beta = 3$. Find the joint pdf of Y_1 and Y_n.

4.3.12. Let $X_1, \ldots, X_n$ be a random sample from a geometric distribution with probability mass function

$$p_i = P(X = i) = pq^{i-1}, i = 1, 2, \ldots, 0 < p < 1, q = 1-p.$$

Show that:

$$P(Y_k = y) = \sum_{i=k}^{n} \binom{n}{i} q^{(y-1)(n-i)}\left\{q^{n-i}[1 - q^y]^i - \left[1 - q^{y-1}\right]^i\right\},$$

$$y = 1, 2, \ldots .$$

4.4 The normal approximation to the binomial distribution

We know that a binomial random variable Y, with parameters n and $P = P(\text{success})$, can be viewed as the number of successes in n trials and can be written as:

$$Y = \sum_{i=1}^{n} X_i$$

where

$$X_i = \begin{cases} 1, & \text{with} \quad \text{probability } p \\ 0, & \text{with} \quad \text{probability} \quad (1-p). \end{cases}$$

The fraction of successes in n trials is:

$$\frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Hence, Y/n is a sample mean. Since $E(X_i) = P$ and $Var\ (X_i) = P(1 - P)$, we have:

$$E\left(\frac{Y}{n}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}np = p$$

and

$$Var\left(\frac{Y}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{p(1-p)}{n}.$$

Because $Y = n\overline{X}$, by the central limit theorem, Y has an approximate normal distribution with mean $\mu = np$ and variance $\sigma^2 = np(1 - P)$. Because the calculation of the binomial probabilities is cumbersome for large sample sizes n, the normal approximation to the binomial distribution is widely used. A useful rule of thumb for use of the normal

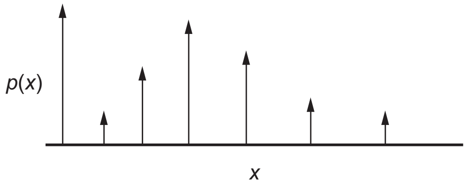


FIGURE 4.7 Probability function of discrete random variable *X*.

approximation to the binomial distribution is to make sure *n* is large enough if $np \geq 5$ and $n(1 - P) \geq 5$. Otherwise, the binomial distribution may be so asymmetric that the normal distribution may not provide a good approximation. Other rules, such as $np \geq 10$ and $n(1 - P) \geq 10$, or $np(1 - P) \geq 10$, are also used in the literature. Because all of these rules are only approximations, for consistency's sake we will use $np \geq 5$ and $n(1 - P) \geq 5$ to test for largeness of sample size in the normal approximation to the binomial distribution. If the need arises, we could use the more stringent condition $np(1 - P) \geq 10$.

Recall that discrete random variables take no values between integers, and their probabilities are concentrated at the integers as shown in Fig. 4.7. However, the normal random variables have zero probability at these integers; they have nonzero probability only over intervals. Because we are approximating a discrete distribution with a continuous distribution, we need to introduce a correction factor for continuity which is explained next.

Correction for continuity for the normal approximation to the binomial distribution

(a) To approximate $P(X \leq a)$ or $P(X > a)$, the correction for continuity is $(a + 0.5)$, that is,

$$P(X \leq a) = P\left(Z < \frac{(a + 0.5) - np}{\sqrt{np(1 - p)}}\right)$$

and

$$P(X > a) = P\left(Z > \frac{(a + 0.5) - np}{\sqrt{np(1 - p)}}\right).$$

(b) To approximate $P(X \geq a)$ or $P(X < a)$, the correction for continuity is $(a - 0.5)$, that is,

$$P(X \geq a) = P\left(Z > \frac{(a - 0.5) - np}{\sqrt{np(1 - p)}}\right)$$

and

$$P(X < a) = P\left(Z < \frac{(a - 0.5) - np}{\sqrt{np(1 - p)}}\right).$$

(c) To approximate $P(a \leq X \leq b)$, treat ends of the intervals separately, calculating two distinct *z*-values according to steps (a) and (b), that is,

$$P(a \leq X \leq b) = P\left(\frac{(a - 0.5) - np}{\sqrt{np(1 - p)}} < Z < \frac{(b + 0.5) - np}{\sqrt{np(1 - p)}}\right).$$

(d) Use the normal table to obtain the approximate probability of the binomial event.

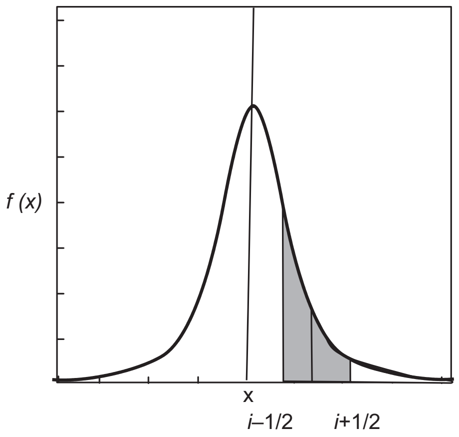


FIGURE 4.8 Continuity correction for $P(X = i)$.

The shaded area in Fig. 4.8 represents the continuity correction for $P(X = i)$.

EXAMPLE 4.4.2

A study of parallel interchange ramps revealed that many drivers do not use the entire length of parallel lanes for acceleration, but seek, as soon as possible, a gap in the major stream of traffic to merge. At one site on Interstate Highway 75, 46% of drivers used less than one-third of the lane length available before merging. Suppose we monitor the merging pattern of a random sample of 250 drivers at this site.

(a) What is the probability that fewer than 120 of the drivers will use less than one-third of the acceleration lane length before merging?

(b) What is the probability that more than 225 of the drivers will use less than one-third of the acceleration lane length before merging?

Solution

First we check for adequacy of the sample size:

$$np = (250)(0.46) = 115 \quad \text{and} \quad n(1-p) = (250)(1-0.46) = 135.$$

Both are greater than 5. Hence, we can use the normal approximation. Let X *be the number of drivers using less than one-third of the lane length available before merging. Then* X *can be considered to be a binomial random variable. Also,*

$$\mu = np = (250)(0.46) = 115.0$$

and

$$\sigma = \sqrt{np(1-p)} = \sqrt{250(0.46)(0.54)} = 7.8804.$$

Thus,

(a) $P(X < 120) = P\left(Z < \frac{119.5-115}{7.8804} = 0.57103\right) = 0.7157$, *that is, we are approximately 71.57% certain that fewer than 120 drivers will use less than one-third of the acceleration length before merging.*

(b) $P(X > 225) = P\left(Z > \frac{225.5-115}{7.8804} = 14.02213\right) \approx 0$, *that is, there is almost no chance that more than 225 drivers will use less than one-third of the acceleration lane length before merging.*

Exercises 4.4

4.4.1 Suppose X is a binomial random variable with $n = 20$ and $P = 0.2$. Find the probability that $X \leq 10$ using binomial tables and compare this with the corresponding value found from normal approximation.

4.4.2. Using normal approximation, find the probability of obtaining at least 90 heads in 150 tosses of a fair coin. Is the normal approximation valid? Why?

4.4.3. A car rental company finds that each day 6% of the persons making reservations will not show up. If the rental company reserves for 215 persons with only 200 automobiles, what is the probability that an automobile will be available for every person who shows up holding a reservation? (Use the normal approximation.)

4.4.4. The president of the United States is thought to have a positive approval rating of 58% of the people at a certain time. In a random sample of 1200 people, what is the approximate probability that the number of positive approvals will be at least 750? Interpret your results and state any assumptions.

4.4.5. In the United States, sudden infant death syndrome (SIDS) is one of the leading causes of postneonatal deaths (those occurring between the ages of 28 days and 1 year). Thus far, the most significant risk factor discovered for SIDS is placing babies to sleep in a prone position (on their stomachs). Suppose the rate of death due to SIDS is 0.00103 per year. In a random sample of 5000 infants between the ages of 28 days and 1 year, what is the approximate probability that the number of SIDS-related deaths will be at least 10? Interpret your results and state any assumptions.

4.4.6. Let X and Y be independent binomial random variables with parameters (n, P_1) and (m, P_2), respectively.

(a) Find $E\left(\frac{X}{n} - \frac{Y}{m}\right)$.

(b) Find $Var\left(\frac{X}{n} - \frac{Y}{m}\right)$.

(c) Show that $\left(\frac{X}{n} - \frac{Y}{m}\right) \sim N\left(E\left(\frac{X}{n} - \frac{Y}{m}\right), Var\left(\frac{X}{n} - \frac{Y}{m}\right)\right)$, for large m and n.

4.5 Chapter summary

In this chapter, we learned about sampling distributions. In sampling distributions associated with normal populations, we have seen that we can generate chi-square, t-, and F-distributions. In Section 4.3 we dealt with order statistics. Then in Section 4.4 we looked at large sample approximations such as the normal approximation to the binomial distribution. In the following section, we will give Minitab examples to show how the idea of sampling distribution can be explored using statistical software.

We will now list some of the key definitions introduced in this chapter:

- Sampling distribution
- Sample and sample size
- Random sample
- Statistic
- Standard error
- Finite population correction factor
- Degrees of freedom
- *t* Distribution
- *F*-distribution
- Order statistics

In this chapter, we have also presented the following important concepts and procedures:

- Sampling distribution associated with normal distribution
- Results on chi-square distribution
- Results on Student *t* Distribution
- Results on *F*-distribution
- Derivation of pdfs for order statistics
- Large sample approximations
- Normal approximation to the binomial
- Correction for continuity for the normal approximation to the binomial distribution

4.6 Computer examples

4.6.1 Examples using R

Note: For the following problems you are generating random samples; your answers will vary!

EXAMPLE 4.6.1 Generating normal random samples

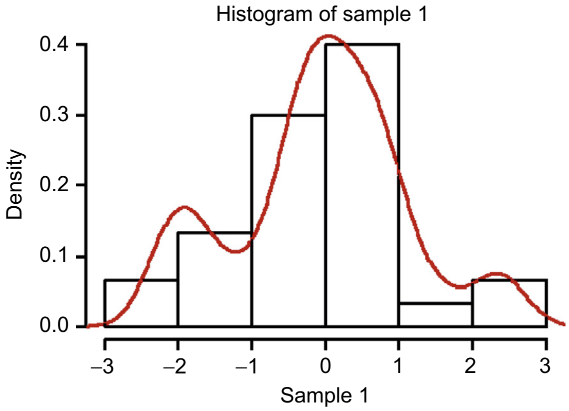
Create three samples of size 30 from standard normal distribution and draw histograms for each sample.

Notice the last two arguments are the mean and standard deviation of the distribution 0, and 1. In addition, plot a density curve over the histogram. Only one output is shown for this example.

R code:

```
sample1 = rnorm(30,0,1);
sample2 = rnorm(30,0,1);
sample3 = rnorm(30,0,1);
hist(sample1,prob = T);
lines(density(sample1),col = “red”);
hist(sample2,prob = T);
lines(density(sample2),col = “red”);
hist(sample3,prob = T);
lines(density(sample3),col = “red”);
```

Output:



EXAMPLE 4.6.2 Generating a normal random sample

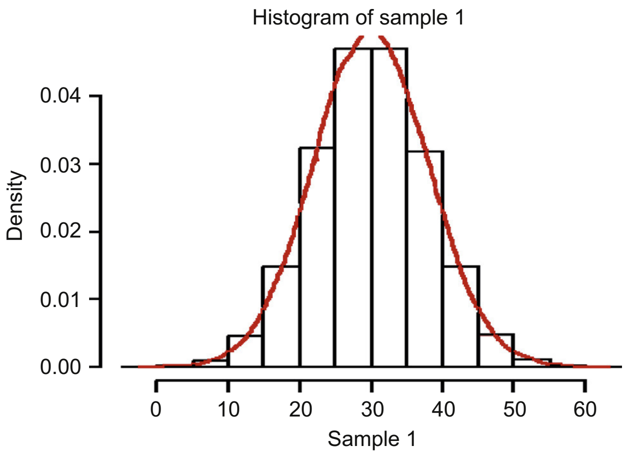
Generate 50,000 observations from a normal distribution with mean 30 and standard deviation 8. Obtain summary statistics for these data and draw a graph.

R code:

```
sample = rnorm(50,000,30,8);
summary(sample);
sd(sample);
hist(sample, prob = T);
lines(density(sample),col = “red”);
```

Output:

```
Min.        1st Qu.     Median      Mean        3rd Qu.     Max.
−0.08056    24.62000    30.01000    30.03000    35.42000    60.82000
7.981,699 ← Standard deviation
```



EXAMPLE 4.6.3 Generating a random exponential sample

From an exponential distribution, draw 10,000 samples, each sample of size 15. Compute the mean of each sample and draw a chart for the means. This will be an approximate sampling distribution of $\overline{x}$ for a fixed sample of size 15.

R code:

```
samples_means = c(); ##Creates an empty array for us to store the means in.
for(i in 1:10,000) { ## This for loop repeats the code inside it change variable i over the range.
  sample = rexp(15,3); ##Generates a random sample of 15 from an exponential.
  mean = mean(sample); ## calculates the mean of that sample.
  samples_means = c(sample, mean); ## store the mean inside our array for later use.
}
hist(samples_means, prob = T); ##Use previous methods to check the distribution of the means.
lines(density(samples_means),col = "red");
summary(samples_means);
sd(samples_means);
```

Output:

No output is given for this particular problem, please see the graph generated by R.

You have stored the samples_means in this variable use previous analysis methods on this variable.

4.6.2 Minitab examples

EXAMPLE 4.6.4

Create three samples of size 30 from standard normal distribution using Minitab, and draw histograms for each sample.

Solution

We can use the following procedure:

1. *Open a new worksheet.*
2. *Choose **Calc** > **Random Data** > **Normal.***
3. *Generate **30** rows of data.*
4. *Store results in **C1−C3.***
5. *Enter a mean of **0** and a standard deviation of **1** and click **OK.***
6. *Choose **Graph** > **Character Graphs** > **Histogram** and enter **C1−C3** in the variable box and click **OK**. We will not give the data or any of the three histograms that we will get. These histograms are just lines containing *s. If we need actual histograms, in step 6 use*

Graph** > **Histogram and enter ***C1*** in the graph variable box and click ***OK***.

If we wish to generate descriptive statistics, then:

7. *Choose **Stat** > **Basic Statistics** > **Display Descriptive statistics** ..., enter **C1−C3** in the variable box, and click **OK.***

If we would like to see the mean for the three samples:

8. *Choose **Calc** > **Row Statistics,** then click **Mean** and in the Input variables type **C1–C3.** In Store Result in: **C4** and click **OK.***
   *To see the histogram of these averages, follow step 6 with **C4** in the graph variable box.*
   Using a similar procedure, one could generate samples from normal distributions with different means and standard deviations, as well as from other distributions.

4.6.3 SPSS examples

If we have the full version of SPSS, we can write code that can be used to simulate a sampling distribution with different values of *P*. However, with the student version, it is not easy to simulate. Therefore, we will not give SPSS examples in this chapter.

4.6.4 SAS examples

EXAMPLE 4.6.5

Generate 50,000 observations from a normal distribution with mean 30 and standard deviation 8. Obtain summary statistics for these data and draw a graph.

Solution

We could use the following program.

```
title '50,000 Obs Sample from a Normal Distribution';
title2 'with Mean = 30 and Standard Deviation = 8';
data normaldat;
  do n = 1 to 50,000;
    X = 8*rannor(55)+30;
    output;
  end;
run;
proc univariate data = normaldat;
  var x;
run;
proc chart;
  vbar x/midpoints = 6 to 54 by 2;
  format x msd.;
run;
```

*In the foregoing program, rannor **(55)**, the number 55 is just a seed number to obtain the same series of random numbers each time we run the program. If we use 0, each time we run the program we will get a different set of random numbers. We will not give the output.*

EXAMPLE 4.6.6

From an exponential distribution, draw 10,000 samples, each sample of size 15. Compute the mean of each sample and draw a chart for the means. This will be an approximate sampling distribution of $\overline{X}$ for a fixed sample of size 15.

Solution

Use the following program.

```
title '10,000 Sample Means with 15 Obs per Sample';
title2 'Drawn from an Exponential Distribution';
data sample15;
  do Sample = 1 to 10,000;
    do n = 1 to 15;
      X = ranexp(3);
      output;
    end;
   end;
  proc means data = sample 15 noprint;
  output out = mean 15 mean = Mean;
  var x;
  by sample;
run;
proc chart data = mean 15;
  vbar mean/axis = 1800.
    midpoints = 0.10 to 2.05 by 0.1;
run;
proc univariate data = mean4 noextrobs = 0 normal.
      mu0 = 1;
  mean;
run;
```

This will produce an approximate sampling distribution of $\overline{X}$. *We will not give the output.*

Projects for chapter 4

4A A method to obtain random samples from different distributions

Most of the statistical software packages contain a random number generator that produces approximations to random numbers from the uniform distribution U [0, 1]. To simulate the observation of any other continuous random variables, we can start with uniform random numbers and associate these with the distribution we want to simulate. For example, suppose we wish to simulate an observation from the exponential distribution:

$$F(x) = 1 - e^{-0.5x}, \quad 0 < x < \infty.$$

First produce the value of y from the uniform distribution. Then solve for x from the equation:

$$y = F(x) = 1 - e^{-0.5x}.$$

So $x = [-\ln (1 - y)]/0.5$ is the corresponding value of the exponential random variable. For instance, if $y = 0.67$, then $x = [-\ln (1 - y)]/0.5 = 2.2173$. If we wish to simulate a sample from the distribution F from the different values of y obtained from the uniform distribution, the procedure is repeated for each new observation x.

(a) Simulate 10 observations of a random variable having exponential distribution with mean and standard deviation both equal to 2.

(b) Select 1500 random samples of size $n = 10$ measurements from a population with an exponential distribution with mean and standard deviation both equal to 2. Calculate the sample mean for each of these 1500 samples and draw a relative frequency histogram. Based on Theorems 4.1.1 and 4.4.1, what can you conclude?

It should be noted that, in general, if $Y \sim U\,(0,\ 1)$ random variable, then we can show that $X = -\frac{\ln Y}{\lambda}$ will give an exponential random variable with parameter λ. Uniform random variable could also be used to generate random variables from other distributions. For example, let U_i' be iid U[0, 1] random variables. Then,

$$X = -2\sum_{i=1}^{v} \ln(U_i) \sim \chi^2_{2v},$$

and

$$Y = -\beta\sum_{i=1}^{\alpha} \ln(U_i) \sim \mathit{Gamma}\ (\alpha, \beta).$$

Of course, these transformations are useful only when v and α are integers. More efficient methods based on Monte Carlo simulations, such as MCMC methods, are discussed in Chapter 13.

4B Simulation experiments

When the derivation via probability rules is too difficult or complicated to be carried out, one can use simulation experiments to obtain information about a statistic's sampling distribution. The following characteristics of the experiment must be specified:

(i) the population distribution (normal with $\mu = 10$ and $\sigma = 2$, exponential with $\lambda = 5$, etc.);
(ii) the sample size n and the statistic of interest ($\overline{X}$, S, etc.);
(iii) the number of replications k (such as $k = 300$).

Then, using a computer program, obtain k different random samples, each of size n, from the designated population distribution. Calculate the value of the statistic for each of the k replications. Construct a histogram for this k statistic. This histogram gives the approximate sampling distribution of the statistic. The larger the value of k, the better will be the approximation.

(a) For your simulation study, use the population distribution as normal with $\mu = 3.4$ and $\sigma = 1.2$.

For $n = 8$ perform $k = 500$ replications and draw a histogram for values of the sample means. Repeat the experiment with $n = 15$, $n = 25$, and $n = 35$ and draw the histograms. Based on this exercise, you will be able to intuitively verify the result that $\overline{X}$ based on a large n tends to be closer to μ than does $\overline{X}$ based on a small n.

(b) Repeat the experiment of (a) with different values of k, such as $k = 200$, $k = 750$, and $k = 1000$.
(c) Repeat the simulation study with different distributions such as exponential distribution.

4C A test for normality

Many statistical procedures require that the population be at least approximately normal. Therefore, a procedure is needed for checking that the sampled data could have come from a normal distribution. There are many procedures, such as the normal-score plot, or Lilliefors test for normality, available in statistics for this purpose. We will describe the *normal-score plot*, which is an effective way to detect deviations from normality. *The normal* scores consist of values of z that divide the axes into equal probability intervals. For a sample of size 4, the normal scores are $-z_{0.20} = -0.84$, $-z_{0.40} = -0.25$, $z_{0.40} = -0.25$, and $z_{0.20} = 0.84$.

Steps to construct a normal plot

1. Rearrange the n data points in ascending order.
2. Obtain the n normal scores.
3. Plot the kth largest observation, versus the kth normal score, for all k.
4. If the data were from a standard normal distribution, the plot would resemble a 45° line through the origin.
5. If the observations were from normal (but not from standard normal), the pattern should still be a straight line. However, the line need not pass through the origin or have a slope 1.
   In applications, a minimum of 15−20 observations is needed to reach a more accurate conclusion.

Exercises

1. For different observations, construct normal plots and check for normality of the corresponding populations.
2. Using software (such as Minitab), generate 15 observations each from the following distributions: (**a**) normal (2, 4), (**b**) uniform (0, 1), (**c**) gamma (2, 4), and (**d**) exponential (2).

   For each of these data sets, draw a probability plot and note the geometry of the plots.