Biostatistics: A Guide to Design, Analysis, and Discovery

Ronald N. Forthofer
Boulder County, Colorado

Eun Sul Lee
Division of Biostatistics
Department of Public Health and Preventive Medicine
Oregon Health & Science University, Portland, Oregon

Mike Hernandez
Department of Biostatistics and Applied Mathematics
M.D. Anderson Cancer Center, Houston, Texas

ELSEVIER
ACADEMIC
PRESS

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Univ. of Washington EX2067
Guardant Health, Inc v. Univ. of Washington
IPR2022-00817

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper.

Copyright © 2007, Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, E-mail: permissions@elsevier.com. You may also complete your request on-line via the Elsevier homepage (http://elsevier.com), by selecting "Support & Contact" then "Copyright and Permission" and then "Obtaining Permissions."

Library of Congress Cataloging-in-Publication Data
Application Submitted

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN 13: 978-0-12-369492-8
ISBN 10: 0-12-369492-2

For information on all Academic Press publications
visit our Web site at www.books.elsevier.com

Printed in the United States of America
06 07 08 09 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

Probability Distributions

5

Chapter Outline

5.1 The Binomial Distribution
5.2 The Poisson Distribution
5.3 The Normal Distribution
5.4 The Central Limit Theorem
5.5 Approximations to the Binomial and Poisson Distributions

This chapter introduces three probability distributions: the binomial and the Poisson for discrete random variables, and the normal for continuous random variables. For a discrete random variable, its probability distribution is a listing of the probabilities of its possible outcomes or a formula for finding the probabilities. For a continuous random variable, its probability distribution is usually expressed as a formula that can be used to find the probability that the variable will fall in a specified interval. Knowledge of the probability distribution (1) allows us to summarize and describe data through the use of a few numbers and (2) helps to place results of experiments in perspective; that is, it allows us to determine whether or not the result is consistent with our ideas. We begin the presentation of probability distributions with the binomial distribution.

5.1 The Binomial Distribution

As its name suggests, the *binomial distribution* refers to random variables with two outcomes. Three examples of random variables with two outcomes are (1) smoking status — a person does or does not smoke, (2) exposure to benzene — a worker was or was not exposed to benzene in the workplace, and (3) health insurance coverage — a person does or does not have health insurance. The random variable of interest in the binomial setting is the number of occurrences of the event under study — for example, the number of adults in a sample of size n who smoke or who have been exposed to benzene or who have health insurance. For the binomial distribution to apply, the status of each subject must be independent of that of the other subjects. For example, in the hypertension question, we are assuming that each person's hypertension status is unaffected by any other person's status.

5.1.1 Binomial Probabilities

We consider a simple example to demonstrate the calculation of binomial probabilities. Suppose that four adults (labeled A, B, C, and D) have been randomly selected and

asked whether or not they currently smoke. The random variable of interest in this example is the number of persons who respond yes to the question about smoking. The possible outcomes of this variable are 0, 1, 2, 3, and 4.

The outcomes (0, 1, 2, 3, and 4) translate to estimates of the proportion of persons who answer yes (0.00, 0.25, 0.50, 0.75, and 1.00, respectively). Any of these outcomes could occur when we draw a random sample of four adults. As a demonstration, let us draw 10 random samples of size 4 from a population in which the proportion of adults who smoke is assumed to be 0.25 (population parameter). We can use a random number table in performing this demonstration. Four 2-digit numbers were taken from the first 10 rows of the first page of random number tables in Appendix B. The 2-digit numbers less 25 are considered smokers. The results are shown here:

Sample	Random Number	No. of Smokers	Prop. Smokers
1	17 17 47 59	2	0.50
2	26 58 06 84	1	0.25
3	24 04 23 38	3	0.75
4	74 83 87 93	0	0.00
5	72 86 25 09	1	0.25
6	82 27 49 45	0	0.00
7	77 58 68 91	0	0.00
8	17 80 21 66	2	0.50
9	10 27 10 61	2	0.50
10	07 78 05 54	2	0.50

Three samples have no smokers (estimate of 0.00); two samples have 1 smoker (0.25); four samples have 2 smokers (0.50); one sample has 3 smokers (0.75); and no sample has 4 smokers (1.00). The sample estimates do not necessarily equal the population parameter, and the estimates can vary considerably. In practice, a single sample is selected, and in making an inference from this one sample to the population, this sample-to-sample variability must be taken into account. The probability distribution does this, as will be seen later. Now let us calculate the binomial probability distribution for a sample size of four.

Suppose that in the population, the proportion of people who would respond "yes" to this question is π, and the probability of a response of "no" is then $1 - \pi$. The probability of each of the outcomes can be found in terms of π by listing all the possible outcomes. Table 5.1 provides this listing.

Since each person is independent of all the other persons, the probability of the joint occurrence of any outcome is simply the product of the probabilities associated with each person's outcome. That is, the probability of 4 yes responses is π^4. In the same way, the probability of three yes responses is $4\pi^3(1 - \pi)$, since there are four occurrences of three yes responses. The probability of two yes responses is $6\pi^2(1 - \pi)^2$; the probability of one yes response is $4\pi(1 - \pi)^3$; and the probability of zero yes responses is $(1 - \pi)^4$. If we know the value of π, we can calculate the numerical value of these probabilities.

Suppose π is 0.25. Then the probability of each outcome is as follows:

$$\Pr\{4 \text{ yes responses}\} = 1 * (0.25)^4 * (0.75)^0 = 0.0039 = \Pr\{0 \text{ no responses}\},$$

$$\Pr\{3 \text{ yes responses}\} = 4 * (0.25)^3 * (0.75)^1 = 0.0469 = \Pr\{1 \text{ no response}\},$$

Table 5.1 Possible binomial outcomes in a sample of size of 4 and their probabilities of occurrence.

Person					
A	**B**	**C**	**D**	**Probability of Occurrence**	
y[a]	y	y	y	$\pi * \pi * \pi * \pi$	$= \pi^4 * (1-\pi)^0$
y	y	y	n	$\pi * \pi * \pi * (1-\pi)$	$= \pi^3 * (1-\pi)^1$
y	y	n	y	$\pi * \pi * (1-\pi) * \pi$	$= \pi^3 * (1-\pi)^1$
y	n	y	y	$\pi * (1-\pi) * \pi * \pi$	$= \pi^3 * (1-\pi)^1$
n	y	y	y	$(1-\pi) * \pi * \pi * \pi$	$= \pi^3 * (1-\pi)^1$
y	y	n	n	$\pi * \pi * (1-\pi) * (1-\pi)$	$= \pi^2 * (1-\pi)^2$
y	n	y	n	$\pi * (1-\pi) * \pi * (1-\pi)$	$= \pi^2 * (1-\pi)^2$
y	n	n	y	$\pi * (1-\pi) * (1-\pi) * \pi$	$= \pi^2 * (1-\pi)^2$
n	y	y	n	$(1-\pi) * \pi * \pi * (1-\pi)$	$= \pi^2 * (1-\pi)^2$
n	y	n	y	$(1-\pi) * \pi * (1-\pi) * \pi$	$= \pi^2 * (1-\pi)^2$
n	n	y	y	$(1-\pi) * (1-\pi) * \pi * \pi$	$= \pi^2 * (1-\pi)^2$
y	n	n	n	$\pi * (1-\pi) * (1-\pi) * (1-\pi)$	$= \pi^1 * (1-\pi)^3$
n	y	n	n	$(1-\pi) * \pi * (1-\pi) * (1-\pi)$	$= \pi^1 * (1-\pi)^3$
n	n	y	n	$(1-\pi) * (1-\pi) * \pi * (1-\pi)$	$= \pi^1 * (1-\pi)^3$
n	n	n	y	$(1-\pi) * (1-\pi) * (1-\pi) * \pi$	$= \pi^1 * (1-\pi)^3$
n	n	n	n	$(1-\pi) * (1-\pi) * (1-\pi) * (1-\pi)$	$= \pi^0 * (1-\pi)^4$

[a]y indicates a yes response and n indicates a no response

$$\Pr\{2 \text{ yes responses}\} = 6 * (0.25)^2 * (0.75)^2 = 0.2109 = \Pr\{2 \text{ no responses}\},$$

$$\Pr\{1 \text{ yes response}\} = 4 * (0.25)^1 * (0.75)^3 = 0.4219 = \Pr\{3 \text{ no responses}\},$$

$$\Pr\{0 \text{ yes responses}\} = 1 * (0.25)^0 * (0.75)^4 = 0.3164 = \Pr\{4 \text{ no responses}\}.$$

The sum of these probabilities is one, as it must be, since these are all the possible outcomes. If the probabilities do not sum to one (with allowance for rounding), a mistake has been made. Figure 5.1 shows a plot of the binomial distribution for n equal to 4 and π equal to 0.25.

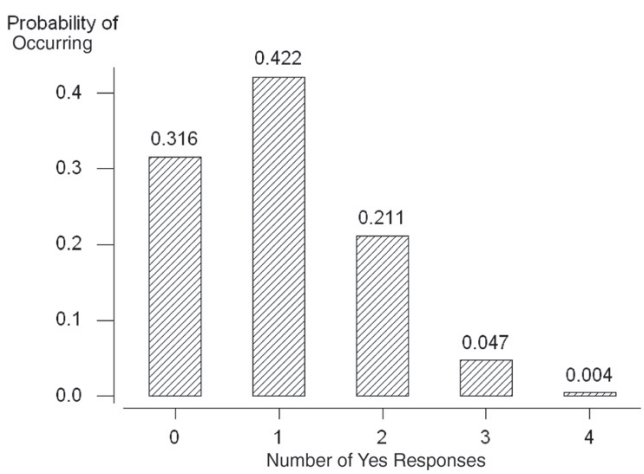

Figure 5.1 Bar chart showing the binomial distribution for $n = 4$ and $\pi = 0.25$.

Are these probabilities reasonable? Since the probability of a yes response is assumed to be 0.25 in the population, in a sample of size four, the probability of one yes response should be the largest. It is also reasonable that the probabilities of zero and two yes

responses are the next largest, since these values are closest to one yes response. The probability of four yes responses is the smallest, as is to be expected. Figure 5.1 shows the rapid decrease in the probabilities as the number of yes responses moves away from the expected response of one.

In the calculation of the probabilities, there are several patterns visible. The exponent of the probability of a yes response matches the number of yes responses being considered and the exponent of the probability of a no response also matches the number of no responses being considered. The sum of the exponents is always the number of persons in the sample. These patterns are easy to capture in a formula, which eliminates the need to enumerate the possible outcomes. The formula may appear complicated, but it is really not all that difficult to use. The formula, also referred to as the *probability mass function* for the binomial distribution, is

$$\Pr\{X = x\} = \binom{n}{x}\pi^x(1-\pi)^{n-x}$$

where $\binom{n}{x} = {}_nC_x = \frac{n!}{x!(n-x)!}$, $n! = n(n-1)(n-2)\cdots 1$ and $0!$ is defined to be 1. The symbol $n!$ is called n factorial, and ${}_nC_x$ is read as n combination x, which gives the number of ways that x elements can be selected from n elements without regard to order (see Appendix A for further explanations). In this formula, n is the number of persons or elements selected, and x is the value of the random variable, which goes from 0 to n. Another representation of this formula is

$$B(x; n, \pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x} = B(n-x; n, 1-\pi)$$

where B represents binomial. The equality of $B(x; n, \pi)$ and $B(n-x; n, 1-\pi)$ is a symbolic way of saying that the probability of x yes responses from n persons, given that π is the probability of a yes response, equals the probability of $n - x$ no responses.

The smoking situation can be used to demonstrate the use of the formula. To find the probability that $X = 3$, we have

$$\Pr(X = 3) = \binom{4}{3}(0.25)^3(0.75)^1 = \frac{4!}{3!1!}(0.015625)(0.75) = 4(0.01172) = 0.0469.$$

This is the same value we found by listing all the outcomes and the associated probabilities. There are easier ways of finding binomial probabilities, as is shown next.

There is a recursive relation between the binomial probabilities, which makes it easier to find them than to use the binomial formula for each different value of X. The relation is

$$\Pr\{X = x+1\} = \left(\frac{n-x}{x+1}\right)\left(\frac{\pi}{1-\pi}\right)\Pr\{X = x\}$$

for x ranging from 0 to $n - 1$. For example, the probability that X equals 1 in terms of the probability that X equals 0 is

$$\Pr\{x=1\}=\left(\frac{4-0}{0+1}\right)\left(\frac{0.25}{0.75}\right)(0.3164)=4\left(\frac{1}{3}\right)(0.3164)=0.4219$$

which is the same value we calculated above.

A still easier method is to use Appendix Table B2, a table of binomial probabilities for n ranging from 2 to 20 and π beginning at 0.01 and ranging from 0.05 to 0.50 in steps of 0.05. There is no need to extend the table to values of π larger than 0.50 because $B(x; n, \pi)$ equals $B(n - x; n, 1 - \pi)$. For example, if π were 0.75 and we wanted to find the probability that $X = 1$ for $n = 4$, $B(1; 4, 0.75)$, we find $B(3; 4, 0.25)$ in Table B2 and read the value of 0.0469. These probabilities are the same because when $n = 4$ and the probability of a yes response is 0.25, the occurrence of three yes responses is the same as the occurrence of one no response when the probability of a no response is 0.75.

Another way of obtaining binomial probabilities is to use computer packages (see **Program Note 5.1** on the website). The use of computer software is particularly nice, since it does not limit the values of π to being a multiple of 0.05 and n can be much larger than 20. More will be said about how large n can be in a later section.

Table 5.2 Probability mass ($\Pr\{X = x\}$) and cumulative ($\Pr\{X \le x\}$) distribution functions for the binomial when $n = 4$ and $\pi = 0.25$.

	x	0	1	2	3	4
Mass:	$\Pr\{X = x\}$	0.3164	0.4219	0.2109	0.0469	0.0039
Cumulative:	$\Pr\{X \le x\}$	0.3164	0.7383	0.9492	0.9961	1.0000

The probability mass function for the binomial gives $\Pr\{X = x\}$ for x ranging from 0 to n (shown in Figure 5.1). Another function that is used frequently is the *cumulative distribution function* (cdf). This function gives the probability that X is less than or equal to x for all possible values of X. Table 5.2 shows both the probability mass function and the cumulative distribution function values for the binomial when n is 4 and π is 0.25. The entries in the cumulative distribution row are simply the sum of the probabilities in the row above it, the probability mass function row, for all values of X less than or equal to the value being considered. Cumulative distribution functions all have a general shape shown in Figure 5.2. The value of the function starts with a low value and then increases over the range of the X variable. The rate of increase of the function is what varies between different distributions. All the distributions eventually reach the value of one or approach it asymptotically.

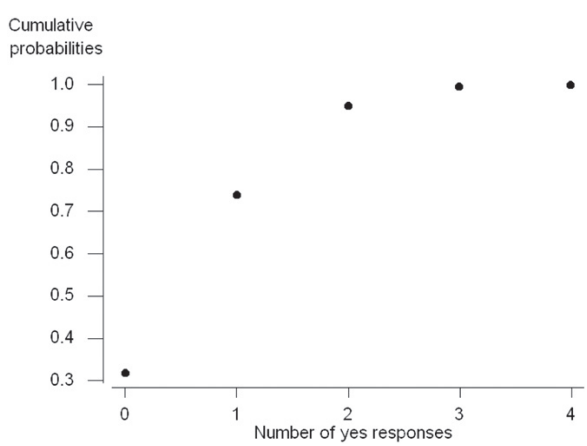

Figure 5.2 Cumulative binomial distribution for $n = 4$ and $\pi = 0.25$.

As seen above, if we know the data follow a binomial distribution, we can completely summarize the data through their two parameters, the sample size and the population proportion or an estimate of it. The sample estimate of the population proportion is the number of occurrences of the event in the sample divided by the sample size.

5.1.2 Mean and Variance of the Binomial Distribution

We can now calculate the *mean* and *variance* of the binomial distribution. The mean is found by summing the product of each outcome by its probability of occurrence — that is,

$$\mu = \sum_{x=0}^{n} x \cdot \Pr\{X = x\}.$$

This appears to be different from the calculation of the sample mean in Chapter 3, but it is really the same because in Chapter 3 all the observations had the same probability of occurrence, $1/N$. Thus, the formula for the population mean could be reexpressed as

$$\sum_{i=1}^{N} x_i \cdot \left(\frac{1}{N}\right) = \sum_{i=1}^{N} x_i \cdot \Pr\{x_i\}.$$

The mean of the binomial variable — that is, the mean number of yes responses out of n responses when n is 4 and π is 0.25 is

$$0 \cdot (0.3164) + 1 \cdot (0.4219) + 2 \cdot (0.2109) + 3 \cdot (0.0469) + 4 \cdot (0.0039) = 1.00 = n\pi$$

or in general for the binomial distribution,

$$\mu = n\pi.$$

The expression of the binomial mean as $n\pi$ makes sense, since, if the probability of occurrence of an event is π, then in a sample of size n, we would expect $n\pi$ occurrences of the event.

The variance of the binomial variable, the number of yes responses, can also be expressed conveniently in terms of π. From Chapter 3, the population variance was expressed as

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)/N.$$

In terms of the binomial, the X variable takes on the values from 0 to n, and we again replace the N in the divisor by the probability that X is equal to x. Thus, the formula becomes

$$\sigma^2 = \sum_{x=0}^{n} (x - n\pi)^2 \Pr\{X = x\}$$

which, with further algebraic manipulation, simplifies to

$$\sigma^2 = n\pi(1 - \pi).$$

When n is 4 and π is 0.25, the variance is then 4(0.25)(1 − 0.25), which is 0.75.

There is often interest in the variance of the proportion of yes responses — that is, in the variance of the number of yes responses divided by the sample size. This is the variance of the number of yes responses divided by a constant. From Chapter 3, we know that this is the variance of the number of yes responses divided by the square of the constant. Thus, the variance of a proportion is

$$\mathrm{Var}(prop.) = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}.$$

Example 5.1

Use of the Binomial Distribution: Let us consider a larger example now. In 1990, cesarean section (c-section) deliveries represented 23.5 percent of all deliveries in the United States, a tremendous increase since 1960 when the rate was only 5.5 percent. Concern has been expressed, for example, by the Public Citizen Health Research Group (1992) in its June 1992 health letter, reporting that many unnecessary c-section deliveries are performed. Public Citizen believes unnecessary c-sections waste resources and increase maternal risks without achieving sufficient concomitant improvement in maternal and infant health. It is in this context that administrators at a local hospital are concerned, as they believe that their hospital's c-section rate is even higher than the national average. Suppose as a first step in determining if this belief is correct, we select a random sample of deliveries from the hospital. Of the 62 delivery records pulled for 1990, we found 22 c-sections. Does this large proportion of c-section deliveries, 35.5 percent (= 22/62), mean that this hospital's rate is higher than the national average? The sample proportion of 35.5 percent is certainly larger than 23.5 percent, but our question refers to the population of deliveries in the hospital in 1990, not the sample. As we just saw, we cannot infer immediately from this sample without taking sample-to-sample variability into account. This is a situation where the binomial distribution can be used to address the question about the population based on the sample.

To put the sample rate into perspective, we need to first answer a question: How likely is a rate of 35.5 percent or higher in our sample if the rate of c-section deliveries is really 23.5 percent? Note that the question includes rates higher than 35.5 percent. We must include them because if the sum of their probabilities is large, we cannot conclude that a rate of 35.5 percent is inconsistent with the national rate regardless of how unlikely the rate of 35.5 percent is.

We can use the cdf for the binomial to find the answer to this question. The cdf enables us to find the probability that a variable is less than a given value — in this case, less than the result we observed in our sample. Then we can subtract that probability from one to find how likely it is to obtain a rate as large or larger than our sample rate. It turns out to be 0.0224 (see **Program Note 5.1** on the website). This means that the probability of 22 or more c-section deliveries is 0.0224. The probability of having 22 or more c-sections is very small. It is unlikely that this hospital's c-section rate is the same as the national average, and, in fact, it appears to be higher. Further investigation is required to determine why the rate may be higher.

5.1.3 Shapes of the Binomial Distribution

The binomial distribution has two parameters, the sample size and the population proportion, that affect its appearance. So far we have seen the distribution of one binomial — Figure 5.1 — which had a sample size of 4 and a population proportion of 0.25. Figure 5.3 examines the effect of population proportion on the shape of the binomial distribution for a sample size of 10.

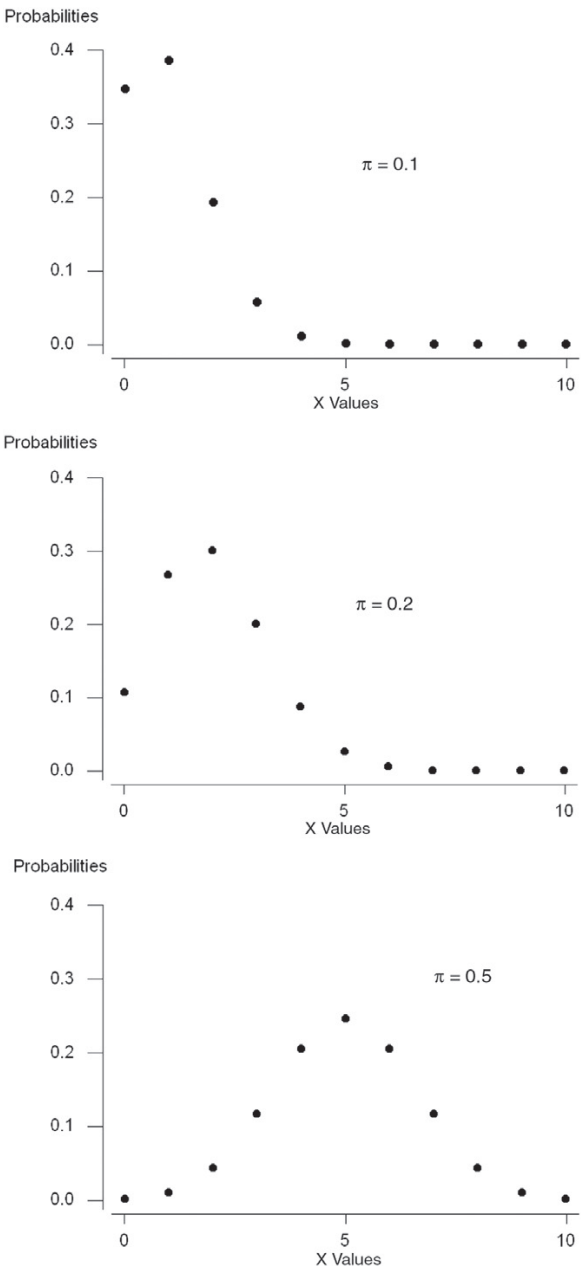

Figure 5.3 Binomial probabilities for $n = 10$ and $\pi = 0.1$, 0.2, and 0.5.

The plots in Figure 5.3 would look like bar charts if a perpendicular line were drawn from the horizontal axis to the points above each outcome. In the first plot with π equal to 0.10, the shape is quite asymmetric with only a few of the outcomes having probabilities very different from zero. This plot has a long tail to the right. In the second plot with π equal to 0.20, the plot is less asymmetric.

The third binomial distribution, with π equal to 0.50, has a mean of 5 ($= n\pi$). The plot is symmetric about its mean of 5, and it has the familiar bell shape. Since π is 0.50, it is as likely to have one less occurrence as one more occurrence — that is, four occurrences of the event of interest are as likely as six occurrences, three as likely as seven and so on, and the plot reflects this.

This completes the introduction to the binomial, although we shall say more about it later. The next section introduces the Poisson distribution, another widely used distribution.

5.2 The Poisson Distribution

The *Poisson distribution* is named for its discoverer, Siméon-Denis Poisson, a French mathematician from the late 18th and early 19th centuries. He is said to have once remarked that life is good for only two things: doing mathematics and teaching it (Boyer 1985). The Poission distribution is similar to the binomial in that it is also used with counts or the number of events. The Poisson is particularly useful when the events occur infrequently. It has been applied in the epidemiologic study of many forms of cancer and other rare diseases over time. It has also been applied to the study of the number of elements in a small space when a large number of these small spaces are spread at random over a much larger space — for example, in the study of bacterial colonies on an agar plate.

Even though the Poisson and binomial distributions both are used with counts, the situations for their applications differ. The binomial is used when a sample of size n is selected and the number of events and nonevents are determined from this sample. The Poisson is used when events occur at random in time or space, and the number of these events is noted. In the Poisson situation, *no* sample of size n has been selected.

5.2.1 Poisson Probabilities

The Poisson distribution arises from either of two models. In one model — quantities, for example — bacteria are assumed to be distributed at random in some medium with a uniform density of λ(lambda) per unit area. The number of bacteria colonies found in a sample area of size A follows the Poisson distribution with a parameter μ equal to the product of λ and A.

In terms of the model over time, we assume that the probability of one event in a short interval of length t_1 is proportional to t_1 — that is, Pr{exactly one event} is approximately λt_1. Another assumption is that t_1 is so short that the probability of more than one event during this interval is almost zero. We also assume that what happens in one time interval is independent of the happenings in another interval. Finally, we assume that λ is constant over time. Given these assumptions, the number of occurrences of the

event in a time interval of length t follows the Poisson distribution with parameter μ, where μ is the product of λ and t.

The Poisson probability mass function is

$$\Pr(X = x) = \frac{e^{-\mu}\mu^x}{x!} \qquad \text{for } x = 0, 1, 2 \ldots\ldots\ldots$$

where e is a constant approximately equal to 2.71828 and μ is the parameter of the Poisson distribution. Usually μ is unknown and we must estimate it from the sample data. Before considering an example, we shall demonstrate in Table 5.3 the use of the probability mass function for the Poisson distribution to calculate the probabilities when $\mu = 1$ and $\mu = 2$. These probabilities are not difficult to calculate, particularly when μ is an integer. There is also a recursive relation between the probability that $X = x + 1$ and the probability that $X = x$ that simplifies the calculations:

$$\Pr\{X = x + 1\} = \left(\frac{\mu}{x+1}\right)\Pr(X = x)$$

for x beginning at a value of 0. For example, for $\mu = 2$,

$$\Pr\{X = 3\} = (2/3)\ \Pr\{X = 2\} = (2/3)\ 0.2707 = 0.1804$$

which is the value shown in Table 5.3.

Table 5.3 Calculation of poisson probabilities, $\Pr\{X = x\} = e^{-\mu}\ \mu^x/x!$, for $\mu = 1$ and 2.

	$\mu = 1$	$\mu = 2$
x	e^{-1} * $1^x/x!$ = $\Pr\{X = x\}$	e^{-2} * $2^x/x!$ = $\Pr\{X = x\}$
0	0.3679 * 1/1 = 0.3679	0.1353 * 1/1 = 0.1353
1	0.3679 * 1/1 = 0.3679	0.1353 * 2/1 = 0.2707
2	0.3679 * 1/2 = 0.1839	0.1353 * 4/2 = 0.2707
3	0.3679 * 1/6 = 0.0613	0.1353 * 8/6 = 0.1804
4	0.3679 * 1/24 = 0.0153	0.1353 * 16/24 = 0.0902
5	0.3679 * 1/120 = 0.0031	0.1353 * 32/120 = 0.0361
6	0.3679 * 1/720 = 0.0005	0.1353 * 64/720 = 0.0120
7	0.3679 * 1/5040 = 0.0001	0.1353 * 128/5040 = 0.0034
8		0.1353 * 256/40320 = 0.0009
9		0.1353 * 512/362880 = 0.0002
	1.0000	0.9999

These probabilities are also found in Appendix Table B3, which gives the Poisson probabilities for values of μ beginning at 0.2 and increasing in increments of 0.2 up to 2.0, then in increments of 0.5 up to 7, and in increments of 1 up to 17. Computer software can provide the Poisson probabilities for other values of μ (see **Program Note 5.1** on the website). Note that the Poisson distribution is totally determined by specifying the value of its one parameter, μ. The plots in Figure 5.4 show the shape of the Poisson probability mass and cumulative distribution functions with $\mu = 2$.

The shape of the Poisson probability mass function with μ equal to 2 (the top plot in Figure 5.4) is similar to the binomial mass function for a sample of size 10 and π equal

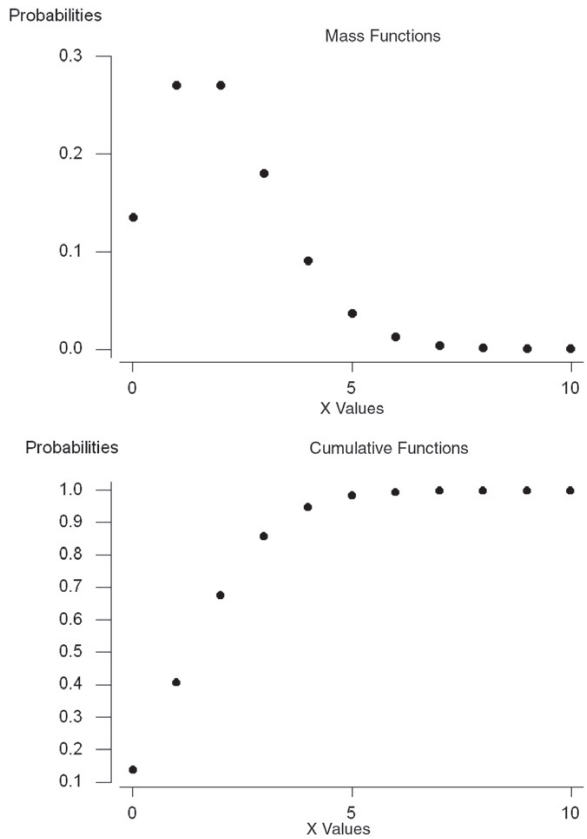

Figure 5.4 Poisson (μ = 2) probability mass and cumulative distribution functions.

to 0.2 just shown. The cdf (the bottom plot in Figure 5.4) has the same general shape as that shown in the preceding binomial example, but the shape is easier to see here, since there are more values for the X variable shown on the horizontal axis.

5.2.2 Mean and Variance of the Poisson Distribution

As just discussed, the mean is found by summing the products of each outcome by its probability of occurrence. For the Poisson distribution with parameter $\mu = 1$ (see Table 5.3), the mean is

$$mean = \sum_{x=0} x \Pr\{X = x\}$$
$$= 0(0.3679) + 1(0.3679) + 2(0.1839) + 3(0.0613)$$
$$+ 4(0.0153) + 5(0.0031) + 6(0.0005) + 7(0.0001) = 1.0000.$$

The mean of the Poisson distribution is the same as μ, which is also the parameter of the Poisson distribution. It turns out that the *variance* of the Poisson distribution is also μ.

5.2.3 Finding Poisson Probabilities

A famous chemist and statistician, W. S. Gosset, worked for the Guinness Brewery in Dublin at the turn of the 20th century. Because Gosset did not wish his competitor breweries to learn of the potential application of his work for a brewery, he published his research under the pseudonym of Student. As part of his work, he studied the distribution of yeast cells over 400 squares of a hemacytometer, an instrument for the counting of cells (Student 1907). One of the four data sets he obtained is shown in Table 5.4.

Table 5.4 Observed frequency of yeast cells in 400 squares.

	X						
	0	**1**	**2**	**3**	**4**	**5**	**6**
Frequency	103	143	98	42	8	4	2
Proportion	0.258	0.358	0.245	0.105	0.020	0.010	0.005
Poisson Probability	0.267	0.352	0.233	0.103	0.034	0.009	0.002

Do these data follow a Poisson distribution? As we just said, the Poisson distribution is determined by the mean value that is unknown in this case. We can use the sample mean to estimate the population mean μ. The sample mean is the sum of all the observations divided by the number of observations — in this case, 400. The sum of the number of cells is

$$103(0) + 143(1) + 98(2) + 42(3) + 8(4) + 4(5) + 2(6) = 529.$$

The sample mean is then 529/400 = 1.3225. Thus, we can calculate the Poisson probabilities using the value of 1.3225 for the mean. Since the value of 1.3225 for μ is not in Appendix Table B3, we must use some other means of obtaining the probabilities. We can calculate them using the recursive relation just shown. We begin by finding the probability of squares with zero cells, $e^{-1.3225}$, which is 0.2665. The other probabilities are found from this value. Computer packages can be used to calculate Poisson probabilities (see **Program Note 5.1** on the website). The results of calculation are shown in the third row of Table 5.4. Based on the visual agreement of the actual and theoretical proportions (from the Poisson), we cannot rule out the Poisson distribution as the distribution of the cell counts. The Poisson distribution agreed quite well for three of the four replications of the 400 cells that Gosset performed.

One reason for interest in the distribution of data is that knowledge of the distribution can be used in future occurrences of this situation. If future data do not follow the previously observed distribution, this can alert us to a change in the process for generating the data. It could also indicate, for example, that the blood cell counts of a patient under study differ from those expected in a healthy population or that there are more occurrences of some disease than was expected assuming that the disease occurrence follows a Poisson distribution with parameter μ. If there are more cases of the disease, it may indicate that there is some common source of infection — for example, some exposure in the workplace or in the environment.

A method of visual inspection of whether or not the data could come from a Poisson distribution is the *Poissonness plot*, presented by Hoaglin (1980). The rationale for the

plot is based on the Poisson probability mass distribution formula. If the data could come from a Poisson distribution, then a plot of the sum of the natural logarithm of the frequency of x and the natural logarithm of $x!$ against the value of x should be a straight line. Using a computer package (see **Program Note 5.2** on the website) with the data in Table 5.4, a Poissonness plot is created, as shown in Figure 5.5.

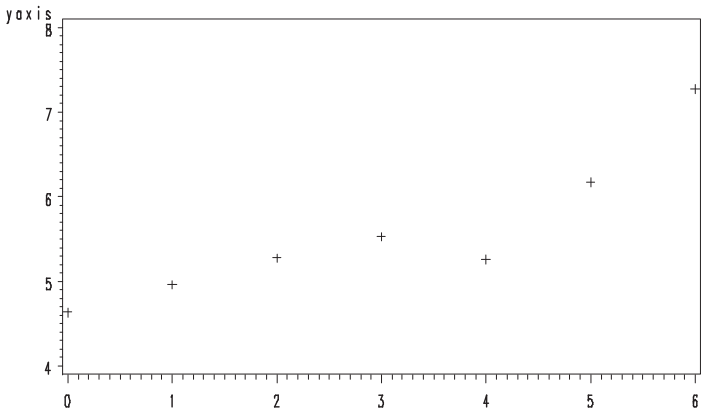

Figure 5.5 Poissonness plot for Gosset's data in Table 5.4.

The plot appears to be approximately a straight line, with the exception of a dip for $x = 4$. In Table 5.4, we see that the biggest discrepancy between the actual and theoretical proportions occurred when $x = 4$, confirmed by the Poissonness plot.

Example 5.2

Use of the Poisson Distribution: In 1986, there were 18 cases of pertussis reported in Harris County, Texas, from its estimated 1986 population of 2,942,550. The reported national rate of pertussis was 1.2 cases per 100,000 population (Harris County Health Department 1990). Do the Harris County data appear to be consistent with the national rate?

The data are inconsistent if there are too many or too few cases of pertussis compared to the national rate. This concern about both too few as well as too many adds a complication lacking in the binomial example in which we were concerned only about too many occurrences. Our method of answering the question is as follows.

First calculate the pertussis rate in Harris County. If the rate is above the national rate, find the probability of at least as many cases occurring as were observed. If the rate is below the national rate, find the probability of the observed number of cases or fewer occurring. To account for both too few as well as too many in our calculations, we double the calculated probability. Is the resultant probability large? If it is large, there is no evidence that the data are inconsistent with the national rate. If it is small, it is unlikely that the data are consistent with the national rate.

The rate of pertussis in Harris County was 0.61 cases per 100,000 population, less than the national rate. Therefore, we shall calculate the probability of 18 or fewer cases given the national rate of 1.2 cases per 100,000 population. The rate of 1.2 per 100,000 is multiplied by 29.4255 (the Harris County population of 2,942,550 divided by 100,000) to obtain the Poisson parameter for Harris County of 35.31. This value exceeds those listed in Table B3. Therefore, we can either find the probability of zero cases and use the recursive formula shown above or use the computer. Using a computer package (see **Program Note 5.1** on the website), the probability of 18 or fewer cases is found to be 0.001. Multiplying this value by 2 to account for the upper tail of the distribution gives a probability of 0.002, a very small value. It is therefore doubtful, since the probability is only 0.002, that the national rate of pertussis applies to Harris County.

This completes the introduction to the binomial and Poisson distributions. The following section introduces the normal probability distribution for continuous random variables.

5.3 The Normal Distribution

The *normal distribution* is also sometimes referred to as the *Gaussian distribution* after the German mathematician Carl Gauss (1777–1855). Gauss, perhaps the greatest mathematician who ever lived, demonstrated the importance of the normal distribution in describing errors in astronomical observations (published in 1809), and today it is the most widely used probability distribution in statistics. Recently, historians discovered that an American mine engineer, Adrian, used the similar distribution for random errors of measurements (published in 1808) (Stigler 1980). The normal distribution is so widely used because (1) it occurs naturally in many situations, (2) the sample means of many nonnormal distributions tend to follow it, and (3) it can serve as a good approximation to some nonnormal distributions.

5.3.1 Normal Probabilities

As we just mentioned, the probability distribution for a continuous random variable is usually expressed as a formula that can be used to find the probability that the continuous variable is within a specified interval. This differs from the probability distribution of a discrete variable that gives the probability of each possible outcome.

One reason why an interval is used with a continuous variable instead of considering each possible outcome is that there is really no interest in each distinct outcome. For example, when someone expresses an interest in knowing the probability that a male 45 to 54 years old weighs 160 pounds, exactly 160.000000000 . . . pounds is not what is intended. What the person intends is related to the precision of the scale used, and the person may actually mean 159.5 to 160.5 pounds. With a less precise scale, 160 pounds may mean a value between 155 and 165 pounds. Hence, the probability distribution of continuous random variables focuses on intervals rather than on exact values.

The probability density function (pdf) for a continuous random variable X is a formula that allows one to find the probability of X being in an interval. Just as the

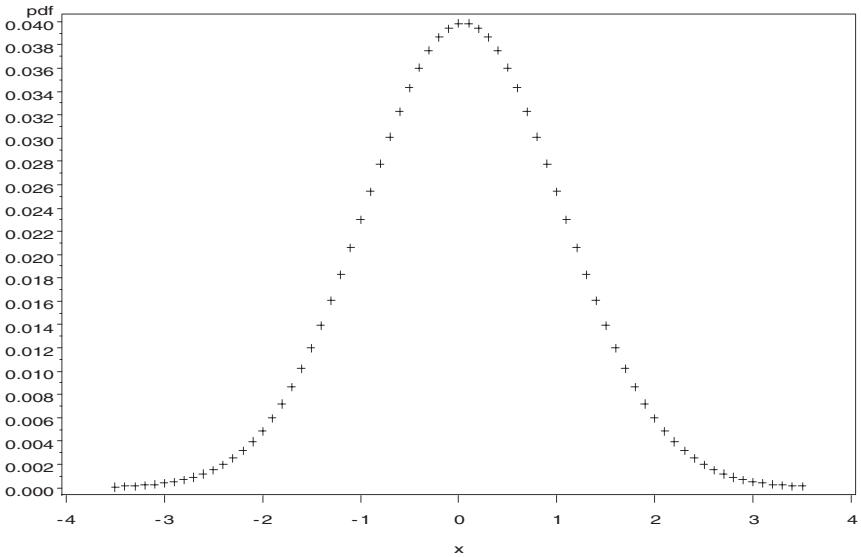

Figure 5.6 Pdf of standard normal distribution.

probability mass function for a discrete random variable could be graphed, the probability density function can also be graphed. Its graph is a curve such that the area under the curve sums to one, and the area between two points, x_1 and x_2, is equal to the probability that the random variable X is between x_1 and x_2.

The normal probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

where μ is the mean and σ is the standard deviation of the normal distribution, and π is a constant approximately equal to 3.14159. The normal probability density function is bell-shaped, as can be seen from Figure 5.6. It shows the standard normal density function — that is, the normal pdf with a mean of zero and a standard deviation of one — over the range of –3.5 to plus 3.5. The area under the curve is one and the probability of X being between any two points is equal to the area under the curve between those two points.

Figure 5.7 shows the effect of increasing σ from one to two on the normal pdf. The area under both of these curves again is one, and both curves are bell-shaped. The standard normal distribution has smaller variability, evidenced by more of the area being closer to zero, as it must, since its standard deviation is 50 percent of that of the other normal distribution. There is more area, or a greater probability of occurrence, under the second curve associated with values farther from the mean of zero than under the standard normal curve. The effect of increasing the standard deviation is to flatten the curve of the pdf, with a concomitant increase in the probability of more extreme values of X.

In Figure 5.8, two additional normal probability density functions are presented to show the effect of changing the mean. Increasing the mean by 3 units has simply shifted the entire pdf curve 3 units to the right. Hence, changing the mean shifts the curve to

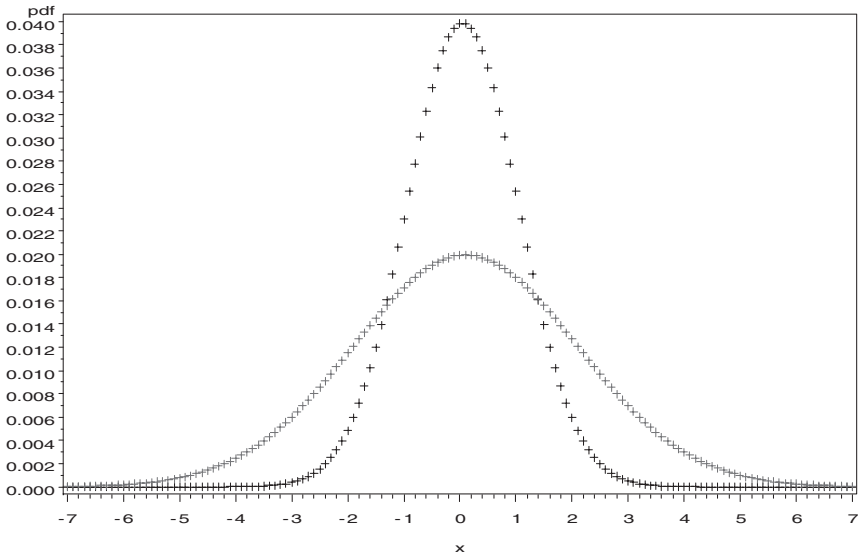

Figure 5.7 Normal pdf of $N(0, 1)$, in black, and $N(0, 2)$, in gray.

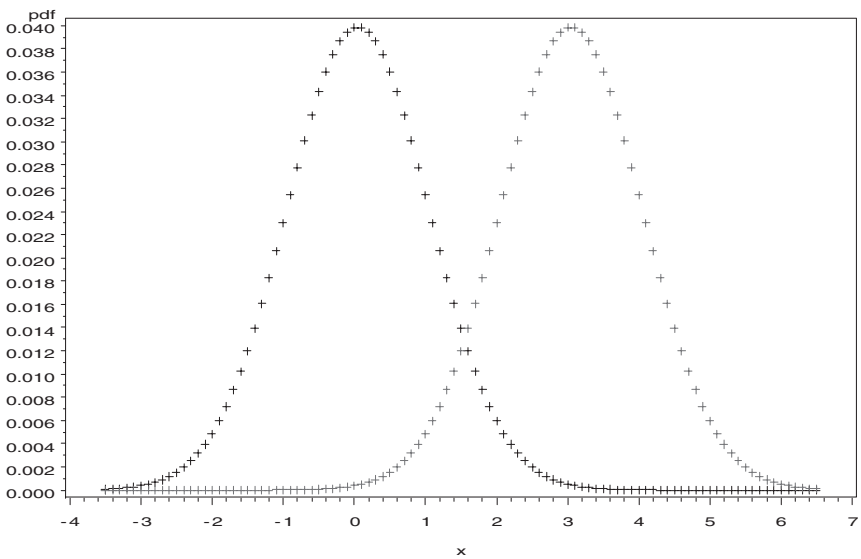

Figure 5.8 Normal pdf of $N(0, 1)$, in black, and $N(3, 1)$, in gray.

the right or left and changing the standard deviation increases or decreases the spread of the distribution.

5.3.2 Transforming to the Standard Normal Distribution

As can be seen from the normal pdf formula and the plots, there are two parameters, the mean and the standard deviation, that determine the location and spread of the normal curve. Hence, there are many normal distributions, just as there are many binomial and Poisson distributions. However, it is not necessary to have many pages of

normal tables for each different normal distribution because all the normal distributions can be transformed to the standard normal distribution. Thus, only one normal table is needed, not many different ones.

Consider data from a normal distribution with a mean of μ and a standard deviation of σ. We wish to transform these data to the standard normal distribution that has a mean of zero and a standard deviation of one. The transformation has two steps. The first step is to subtract the mean, μ, from all the observations. In symbols, let y_i be equal to $(x_i - \mu)$. Then the mean of Y is μ_y, which equals

$$\mu_y = \sum \frac{x_i - \mu}{N} = \frac{\sum x_i - N\mu}{N} = \frac{N\mu - N\mu}{N} = 0.$$

The second step is to divide y_i by its standard deviation. Since we have subtracted a constant from the observations of X, the variance and standard deviation of Y is the same as that of X, as was shown in Chapter 3. That is, the standard deviation of Y is also σ. In symbols, let z_i be equal to y_i/σ. What are the mean and standard deviation of Z? The mean is still zero but the standard deviation of Z is one. This is due to the second property of the variance shown in Chapter 3 — namely, when all the observations are divided by a constant, the standard deviation is also divided by that constant. Therefore, the standard deviation of Z is found by dividing σ, the standard deviation of Y, by the constant σ. The value of this ratio is one.

Therefore, any variable, X, that follows a normal distribution with a mean of μ and a standard deviation of σ can be transformed to the standard normal distribution by subtracting μ from all the observations and dividing all the observed deviations by σ. The variable Z, defined as $(X - \mu)/\sigma$, follows the standard normal distribution. A symbol for indicating that a variable follows a particular distribution or is "distributed as" is the asymptote, ~. For example, $Z \sim N(0, 1)$ means that Z follows a normal distribution with a mean of zero and a standard deviation of one. The observed value of a variable from a standard normal distribution tells how many standard deviations that value is from its mean of zero.

5.3.3 Calculation of Normal Probabilities

The cumulative distribution function of the standard normal distribution, denoted by $\Phi(z)$, represents the probability that the standard normal variable Z is less than or equal to the value z — that is, $\Pr\{Z \le z\}$. Table B4 presents the values of $\Phi(z)$ for values of z ranging from –3.79 to 3.79 in steps of 0.01. The table shows that the value of 0.9999 at $z = 3.79$, meaning that the probability of Z less than 3.79 is practically 1.0000. It also means that the area under the curve of pdf function shown in Figure 5.6 is 1.0000, a requirement for any probability distribution.

Figure 5.9 shows the cumulative distribution function for the standard normal distribution. The vertical axis gives the values of the probabilities corresponding to the values of z shown along the horizontal axis. The curve gradually increases from a probability of 0.0 for values of z around –3 to a probability of 0.5 when z is zero (as marked in Figure 5.9) and on to probabilities close to 1.0 for values of z of 3 or larger. We can calculate various probabilities associated with a normal distribution using its cdf without directly resorting to its pdf.

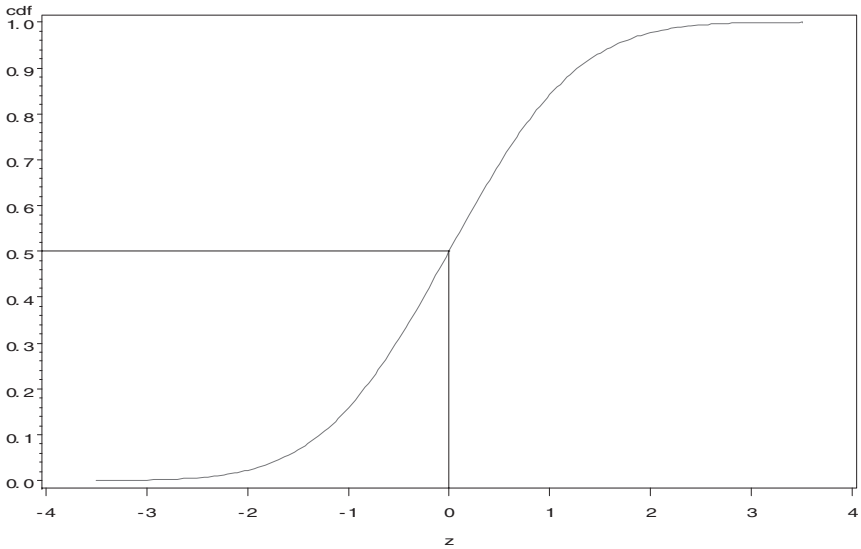

Figure 5.9 Cdf of the standard normal distribution.

Example 5.3

Probability of Being Greater than a Value: Suppose we wish to find the probability that an adult female will have a diastolic blood pressure value greater than 95 mmHg given that X, the diastolic blood pressure for adult females, follows the $N(80, 10)$ distribution. Since the values in Table B4 are for variables that follow the $N(0, 1)$ distribution, we first must transform the value of 95 to its corresponding Z value. To do this, we subtract the mean of 80 and divide by the standard deviation of 10. The z value of 95 mmHg, therefore, is

$$z = \frac{95 - 80}{10} = \frac{15}{10} = 1.5.$$

Thus, the value of the Z variable corresponding to 95 mmHg is 1.5, which means that the diastolic blood pressure of 95 is 1.5 standard deviations above its mean of 80. We now want the probability that Z is greater than 1.5. Using Table B4, look for 1.5 under the z heading and then go across the columns until reaching the .00 column. The probability of a standard normal variable being less than 1.5 is 0.9332. Thus, the probability of being greater than 1.5 is 0.0668 (= 1 − 0.9332).

Example 5.4

Calculation of the Value of the ith Percentile: Table B4 can be used to answer a slightly different question as well. Suppose that we wish to find the 95th percentile of the diastolic blood pressure variable for adult females — that is, the value such that 95 percent of adult females had a diastolic blood pressure less than it. We look in the body of the table until we find 0.9500. We find the corresponding value in the z column, and transform that value to the $N(80, 10)$ distribution. Examination of

Table B4 shows the value of 0.9495 when z is 1.64 and 0.9505 for a z of 1.65. There is no value of 0.9500 in the table. Since 0.9500 is exactly half way between 0.9495 and 0.9505, we shall use the value of 1.645 for the corresponding z. We now must transform this value to the $N(80, 10)$ distribution. This is easy to do since we know the relation between Z and X.

As $Z = (X - \mu)/\sigma$, on multiplication of both sides of the equation by σ, we have $\sigma Z = X - \mu$. If we add μ to both sides of the equation, we have $\sigma Z + \mu = X$. Therefore, we must multiply the value of 1.645 by 10, the value of σ, and add 80, the value of μ, to it to find the value of the 95th percentile. This value is 96.45 (= 16.45 + 80) mmHg.

This calculation can also be performed by computer packages (see **Program Note 5.3** on the website).

The percentiles of the standard normal distribution are used frequently, and, therefore, a shorthand notation for them has been developed. The ith percentile for the standard normal distribution is written as z_i — for example, $z_{0.95}$ is 1.645. From Table B4, we also see that $z_{0.90}$ is approximately 1.28 and $z_{0.975}$ is 1.96. By the symmetry of the normal distribution, we also know that $z_{0.10}$ is −1.28, $z_{0.05}$ is −1.645 and $z_{0.025}$ is −1.96.

The percentiles in theory could also be obtained from the graph of the cdf for the standard normal shown in Figure 5.9. For example, if the 90th percentile was desired, find the value of 0.90 on the vertical axis and draw a line parallel to the horizontal axis from it to the graph. Next, drop a line parallel to the vertical axis from that point down to the horizontal axis. The point where the line intersects the horizontal axis is the 90th percentile of the standard normal distribution.

Example 5.5

Probability Calculation for an Interval: Suppose that we wished to find the proportion of women whose diastolic blood pressure was between 75 and 90 mmHg. The first step in finding the proportion of women whose diastolic blood pressure is in this interval is to convert the values of 75 and 90 mmHg to the $N(0, 1)$ distribution. The value of 75 is transformed to an $N(0, 1)$ value by subtracting μ and dividing by σ — that is, $(75 - 80)/10$, which is −0.5, and 90 is converted to 1.0. We therefore must find the area under the standard normal curve between −0.5 and 1.0. Figure 5.10 aids our understanding of what is wanted. It also provides us with an idea of the probability's value. If the numerical value is not consistent with our idea of the value, perhaps we misused Appendix Table B4. From Figure 5.10 the area under the curve between $z = -0.5$ and $z = 1.0$ appears to be roughly $\frac{1}{2}$ of the total area.

One way of finding the area between −0.5 and 1.0 is to find the area under the curve less than or equal to 1.0 and to subtract from it the area under the curve less than or equal to −0.5. In symbols, this is

$$\Pr\{-0.5 < Z < 1.0\} = \Pr\{Z < 1.0\} - \Pr\{Z < -0.5\}.$$

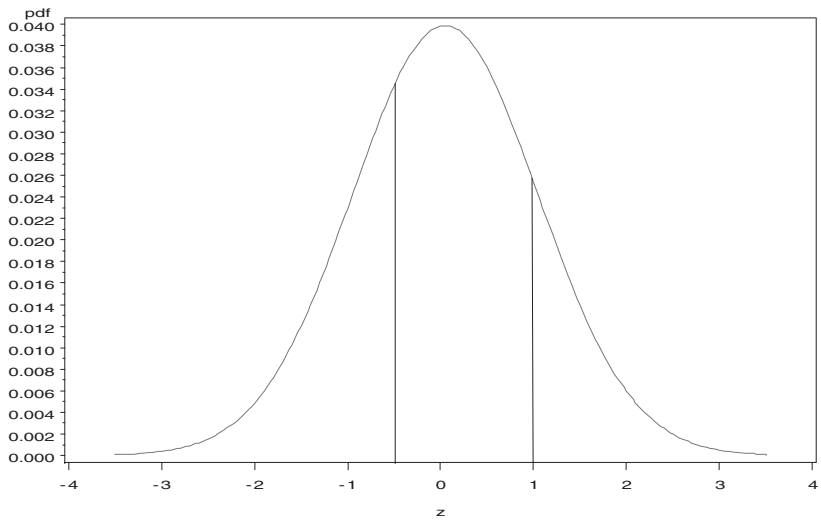

Figure 5.10 Area under the standard normal curve between $z = -0.5$ and $z = 1.0$.

From Table B4, we find that the area under the standard normal pdf curve less than or equal to 1.0 is 0.8413. The probability of a value less than or equal to −0.5 is 0.3085. Thus, the proportion of women whose diastolic blood pressure is between 75 and 90 mmHg is 0.5328 (= 0.8413 − 0.3085). Computer packages can be used to perform this calculation (see **Program Note 5.3** on the website).

5.3.4 The Normal Probability Plot

The normal probability plot provides a way of visually determining whether or not data might be normally distributed. This plot is based on the cdf of the standard normal distribution. Special graph paper, called normal probability paper, is used in the plotting of the points. The vertical axis of normal probability paper shows the values of the cdf of the standard normal. Table B4 shows the cdf values corresponding to z values of –3.79 to 3.79 in steps of 0.01, and it is not difficult to discover that that the increase in the cdf's value is not constant per a constant increase in z. It is more clearly shown in Figure 5.9. The vertical axis reflects this with very small changes in values of the cdf initially, then larger changes in the cdf's values in the middle of plot, and finally very small changes in the cdf's value. Numbers along the horizontal axis are in their natural units.

If a variable, X, is normally distributed, the plot of its cdf against X should be a straight line on normal probability paper. If the plot is not a straight line, then it suggests that X is not normally distributed. Since we do not know the distribution of X, we approximate its cdf in the following fashion.

We first sort the observed values of X from lowest to highest. Next we assign ranks to the observations from 1 for the lowest to n (the sample size) for the highest value. The ranks are divided by n and this gives an estimate of the cdf. This sample estimate is often called the *empirical distribution function*.

The points, determined by the values of the sample estimate of the cdf and the corresponding values of x, are plotted on normal probability paper. In practice, the ranks

divided by the sample size are not used as the estimate of the cdf. Instead, the transformation, (rank – 0.375)/(n + 0.25), is frequently used. One reason for this transformation is that the estimate of the cdf for the largest observation is now a value less than one, whereas the use of the ranks divided by n always results in a sample cdf value of one for the largest observation. A value less than one is desirable because it is highly unlikely that the selected sample actually contains the largest value in the population.

Example 5.6

We consider a small data set for vitamin A values from 33 boys shown in Table 5.5 and examine whether the data are normally distributed. An alternative to normal probability paper is the use of a computer (see **Program Note 5.4** on the website). Applying the probability plot option in a computer package to vitamin A data, Figure 5.11 is produced. The straight line helps to discern whether or not the data deviate from the normal distribution. The points in the plot do not appear to fall along a straight line. Therefore, it is doubtful that the vitamin A variable follows a normal distribution, a conclusion that we had previously reached in the discussion of symmetry in Chapter 3.

Table 5.5 Values of vitamin A, their ranks, and transformed ranks, n = 33.

Vit. A (IUs)	Rank	Trans.[a] Rank	Vit. A (IUs)	Rank	Trans. Rank	Vit. A (IUs)	Rank	Trans. Rank
820	1	0.0188	3747	12	0.3496	6754	23	0.6805
964	2	0.0489	4248	13	0.3797	6761	24	0.7105
1379	3	0.0789	4288	14	0.4098	8034	25	0.7406
1459	4	0.1090	4315	15	0.4398	8516	26	0.7707
1704	5	0.1391	4450	16	0.4699	8631	27	0.8008
1826	6	0.1692	4535	17	0.5000	8675	28	0.8308
1921	7	0.1992	4876	18	0.5301	9490	29	0.8609
2246	8	0.2293	5242	19	0.5602	9710	30	0.8910
2284	9	0.2594	5703	20	0.5902	10451	31	0.9211
2671	10	0.2895	5874	21	0.6203	12493	32	0.9511
2687	11	0.3195	6202	22	0.6504	12812	33	0.9812

Source: From dietary records of 33 boys[7]
[a]Transformed by (rank – 0.375)/(n + 0.25)

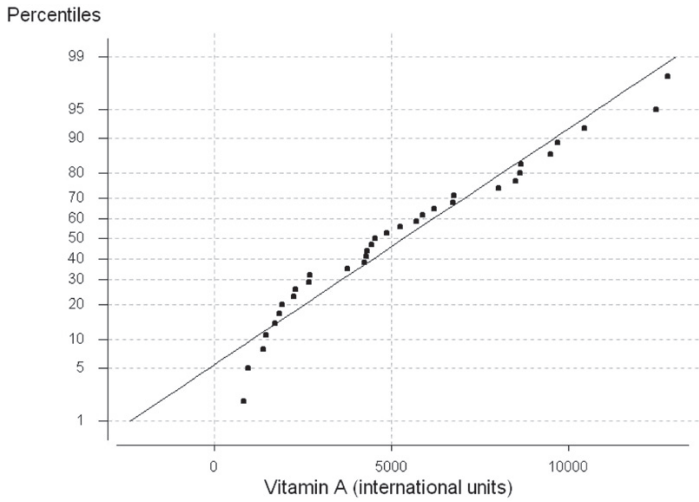

Figure 5.11 Normal probability plot of vitamin A.

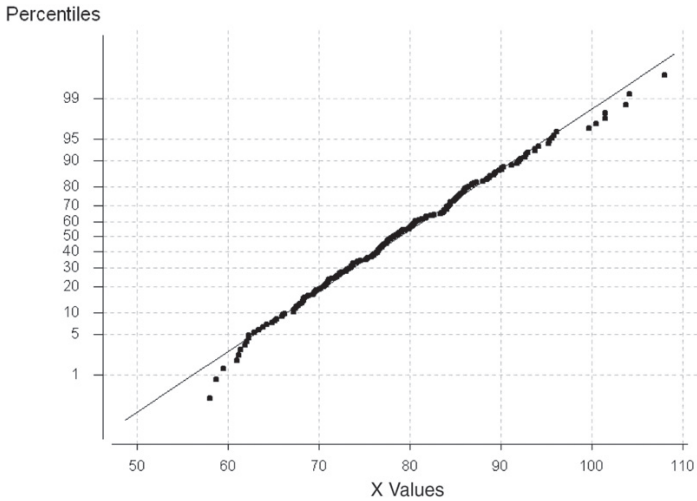

Figure 5.12 Probability plot of 200 observations from N (80, 10).

Let us now examine data from a normal distribution and see what its normality probability plot looks like. The example in Figure 5.12 uses 200 observations generated from an $N(80, 10)$ distribution. The plot looks like a straight line, but there are many points with the same normal scores. The points appear to fall mostly on a straight line as they should. The smallest observed value of X is slightly larger than expected if the data were perfectly normally distributed, but this deviation is relatively slight. Hence, based on this visual inspection, these data could come from a normal distribution.

It is difficult to determine visually whether or not data follow a normal distribution for small sample sizes unless the data deviate substantially from a normal distribution. As the sample size increases, one can have more confidence in the visual determination.

5.4 The Central Limit Theorem

As was just mentioned, one of the main reasons for the widespread use of the normal distribution is that the sample means of many nonnormal distributions tend to follow the normal distribution as the sample size increases. The formal statement of this is called the *central limit theorem*. Basically, for random samples of size n from some distribution with mean μ and standard deviation σ, the distribution of $\bar{x}$, the sample mean, is approximately $N(\mu, \sigma/\sqrt{n})$. This theorem applies for any distribution as long as μ and σ are defined. The approximation to normality improves as n increases.

The proof of this theorem is beyond the scope of this book and also unnecessary for our understanding. We shall, however, demonstrate that it holds for a very nonnormal distribution, the Poisson distribution with mean one.

Example 5.7

As seen Figure 5.4, the Poisson distribution with a mean of 1 is very nonnormal in appearance. The following demonstration consists of drawing a large number of samples — say, 100 — from this distribution, calculating the mean for each sample, and examining the sampling distribution of the sample means. We shall do this for samples of size 5, 10, and 20. Figure 5.13 shows three boxplots for each of these sample sizes. All three means are around 1, and the variances of the means are decreasing as the sample size increases.

As was just stated, the mean of the means should be 1, and the standard deviation of the means is the standard deviation divided by the square root of the sample size. It was also stated that the distribution of means should approach a normal distribution when the sample size is large. Figure 5.14 examines the case for $n = 20$. The mean is 1.003, which is very close to 1. The standard deviation is 0.2058, which is close to $0.2236(=1/\sqrt{20})$. The probability plot lines up around the straight line, suggesting that the distribution of the sample means does not differ substantially from normal distribution.

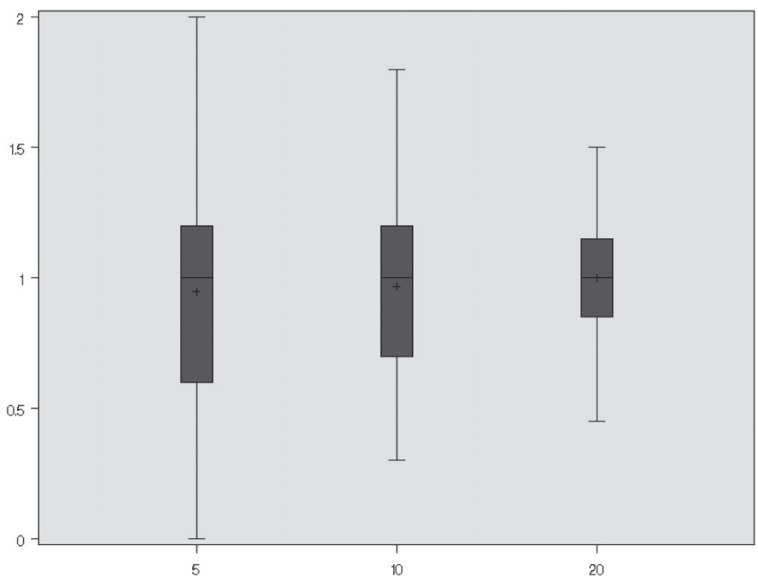

Figure 5.13 Boxplot of 100 sample means from Poisson ($\mu = 1$) for $n = 5$, 10, and 20.

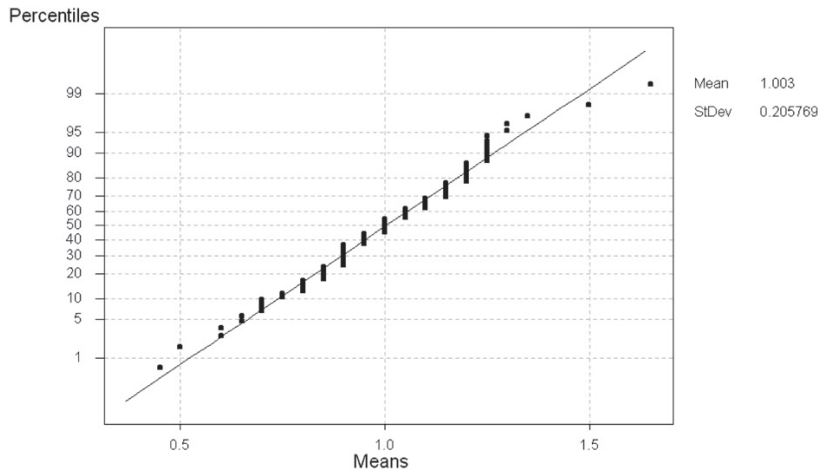

Figure 5.14 Probability plot of 100 means of size 20 from Poisson ($\mu = 1$).

Besides showing that the central limit theorem holds for one very nonnormal distributions, this demonstration also showed the effect of sample size on the estimate of the population mean. This example reinforces the idea that the mean from a very small sample may not be close to the population mean and does not warrant the use of the normal distribution. The idea of central limit theorem and sampling distribution plays a key role in referring from the sample to the population as will be discussed in subsequent chapters.

5.5 Approximations to the Binomial and Poisson Distributions

As we just said, another reason for the use of the normal distribution is that, under certain conditions, it provides a good approximation to some other distributions — in particular the binomial and Poisson distributions. This was more important in the past when there was not such a widespread availability of computer packages for calculating binomial and Poisson probabilities for parameter values far exceeding those shown in tables in most textbooks. However, it is still important today as computer packages have limitations in their ability to calculate binomial probabilities for large sample sizes or for extremely large values of the Poisson parameter. In the following sections, we show the use of the normal distribution as an approximation to the binomial and Poisson distributions.

5.5.1 Normal Approximation to the Binomial Distribution

In the plots of the binomial probability mass functions, we saw that as the binomial proportion approached 0.5, the plot began to look like the normal distribution (see Figure 5.3). This was true for sample sizes even as small as 10. Therefore, it is not surprising that the normal distribution can sometimes serve as a good approximation to the bino-

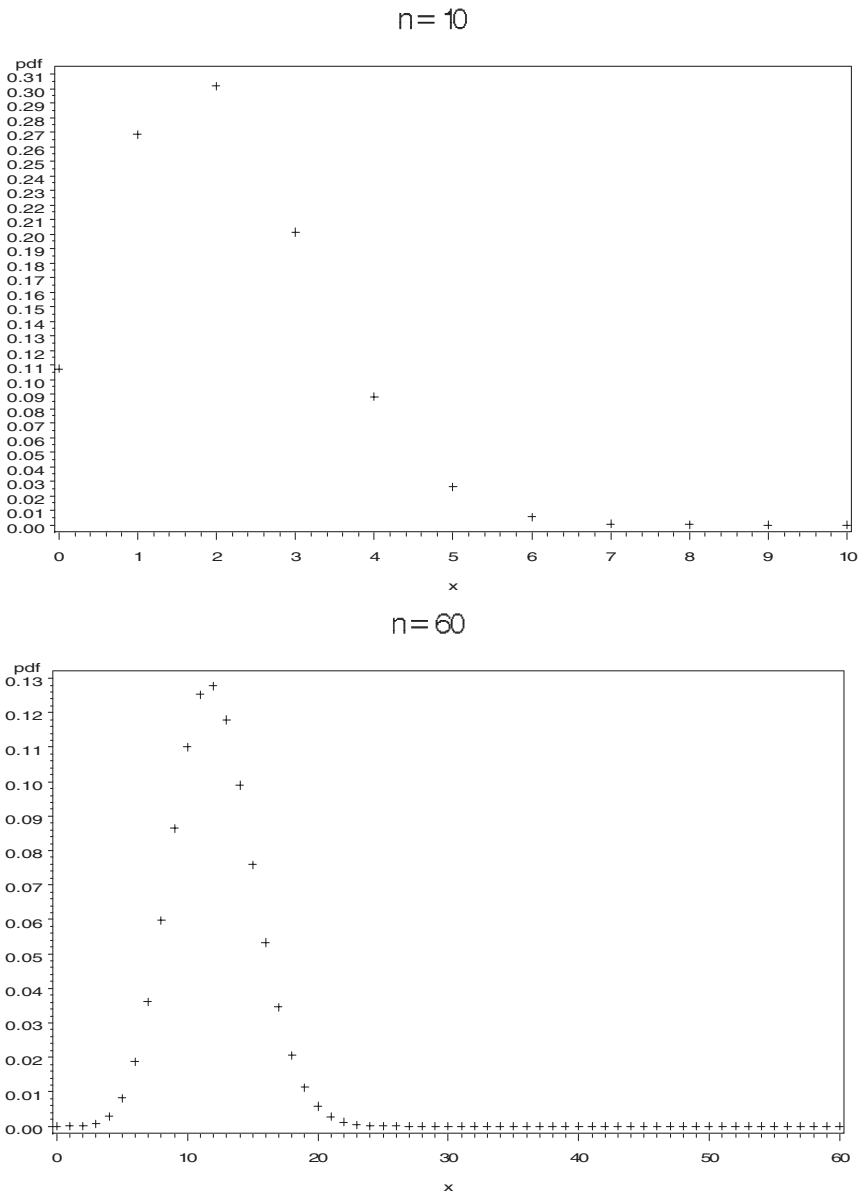

Figure 5.15 Binomial mass functions for $\pi = 0.2$ when $n = 10$ and $n = 60$.

mial distribution. Figure 5.15 demonstrates the effect of n on a binomial distribution, suggesting why we used the modifier *sometimes* in the preceding sentence.

Both plots in Figure 5.15 are based on $\pi = 0.2$. The first plot for $n = 10$ is skewed, and the normal approximation is not warranted. But the second plot for $n = 60$ is symmetric, and the normal distribution should provide a reasonable approximation here.

The central limit theorem provides a rationale for why the normal distribution can provide a good approximation to the binomial. In the binomial setting, there are two

Table 5.6 Sample size required for the normal distribution to serve as a good approximation to the binomial distribution as a function of the binomial proportion π.

π	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
n	440	180	100	60	43	32	23	15	11	10
Difference[a]	0.0041	0.0048	0.0054	0.0059	0.0059	0.0057	0.0059	0.0060	0.0049	0.0027
Mean diff.[b]	0.0010	0.0012	0.0013	0.0016	0.0016	0.0016	0.0016	0.0017	0.0016	0.0013

[a]Maximum difference between binomial probability and normal approximation
[b]Mean of absolute value of difference between binomial probability and normal approximation for all nonzero probabilities

outcomes — for example, disease and no disease. Let us assign the numbers 1 and 0 to the outcomes of disease and no disease, respectively. The sum of these numbers over the entire sample is the number of diseased persons in the sample. The mean, then, is simply the number of diseased sample persons divided by the sample size. And according to the central limit theorem, the sample mean should approximately follow a normal distribution as n increases. But if the sum of values divided by a constant approximately follows a normal distribution, the sum of the values itself also approximately follows a normal distribution. The sum of the values in this case is the binomial variable, and, hence, it also approximately follows the normal distribution.

Unfortunately, there is not a consensus as to when the normal approximation can be used — that is, when n is large enough for the central limit theorem to apply. This issue has been examined in a number of recent articles (Blyth and Still 1983; Samuels and Lu 1992; Schader and Schmid 1989). Based on work by Samuels and Lu (1992) and on some calculations we performed, Table 5.6 shows our recommendations for the size of the sample required as a function of π for the normal distribution to serve as a good approximation to the binomial distribution. Use of these sample sizes guarantees that the maximum difference between the binomial probability and its normal approximation is less than or equal to 0.0060 and that the average difference is less than 0.0017.

The mean and variance to be used in the normal approximation to the binomial are the mean and variance of the binomial, $n\pi$ and $n\pi(1 - \pi)$, respectively. Since we are using a continuous distribution to approximate a discrete distribution, we have to take this into account. We do this by using an interval to represent the integer. For example, the interval of 5.5 to 6.5 would be used with the continuous variable in place of the discrete variable value of 6. This adjustment is called the *correction for continuity.*

Example 5.8

We use the normal approximation to the binomial for the c-section deliveries example in Example 5.1. We wanted to find the probability of 22 or more c-section deliveries in a sample of 62 deliveries. The values of the binomial mean and variance, assuming that π is 0.235, are 14.57 (= 62 ∗ 0.235) and 11.146 (= 62 ∗ 0.235 ∗ 0.765), respectively. The standard deviation of the binomial is then 3.339. Finding the probability of 22 or more c-sections for the discrete binomial variable is approximately equivalent to finding the probability that a normal variable with a mean of 14.57 and a standard deviation of 3.339 is greater than 21.5.

Before using the normal approximation, we must first check to see if the sample size of 62 is large enough. From Table 5.6, we see that since the assumed value of π

is between 0.20 and 0.25, our sample size is large enough. Therefore, it is okay to use the normal approximation to the binomial.

To find the probability of being greater than 21.5, we convert 21.5 to a standard normal value by subtracting the mean and dividing by the standard deviation. The corresponding z value is 2.075 (= [21.5 − 14.57]/3.339). Looking in Table B4, we find the probability of a standard normal variable being less than 2.075 is about 0.9810. Subtracting this value from one gives the value of 0.0190, very close to the exact binomial value of 0.0224 found in Example 5.1.

Example 5.9

According to data reported in Table 65 of *Health*, *United States*, *1991* (NCHS 1992), 14.0% of high school seniors admitted that they used marijuana during the 30 days previous to a survey conducted in 1990. If this percentage applies to all seniors in high school, what is the probability that in a survey of 140 seniors, the number reporting use of marijuana will be between 15 and 25? We want to use the normal approximation to the binomial, but we must first check our sample size with Table 5.7. Since a sample of size 100 is required for a binomial proportion of 0.15, our sample of 140 for an assumed binomial proportion of 0.14 is large enough to use the normal approximation.

The mean of the binomial is 19.6 and the variance is 16.856 (= 140 ∗ 0.14 ∗ 0.86). Thus, the standard deviation is 4.106. These values are used in converting the values of 15 and 25 to z scores. Taking the continuity correction into account means that interval is really from 14.5 to 25.5.

We convert 14.5 and 25.5 to z scores by subtracting the mean of 19.6 and dividing by the standard deviation of 4.106. The z scores are −1.24 (= [14.5 − 19.6]/4.106) and 1.44(= [25.5 − 19.6]/4.106). To find the probability of being between −1.24 and 1.44, we will first find the probability of being less than 1.44. From that, we will subtract the probability of being less than −1.24. This subtraction yields the probability of being in the interval.

These probabilities are found from Table B4 in the following manner. First, we read down the z column until we find the value of 1.44. We go across to the .00 column and read the value of 0.9251; this is the probability of a standard normal value being less than 1.44. The probability of being less than −1.24 is 0.1075. Subtracting 0.1075 from 0.9251 yields 0.8176. This is the probability that, out of a sample of 140, between 15 to 25 high school seniors would admit to using marijuana during the 30 days previous to the question being asked.

5.5.2 Normal Approximation to the Poisson Distribution

Since the Poisson tables do not show every possible value of the parameter μ, and since the tables and computer packages do not provide probabilities for extremely large values of μ, it is useful to be able to approximate the Poisson distribution. As can be seen from the preceding plots, the Poisson distribution does not look like a normal distribution for

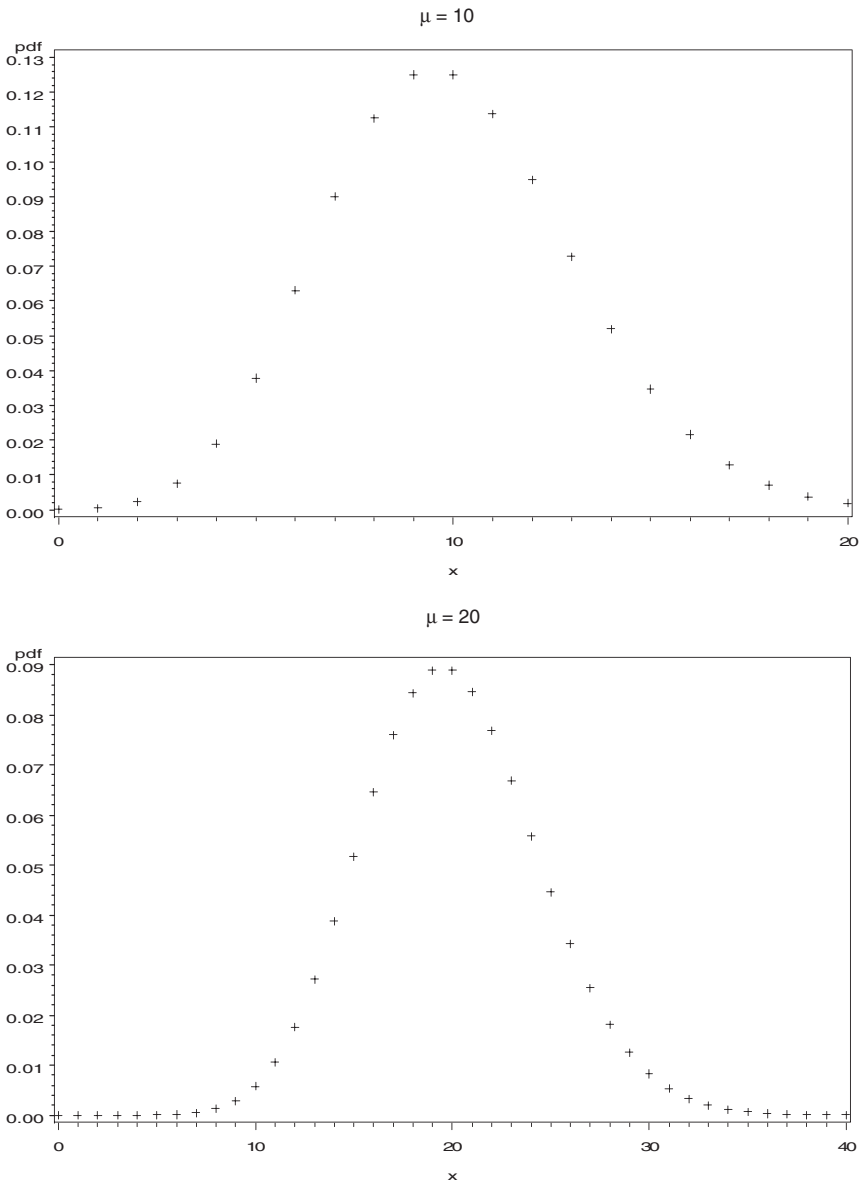

Figure 5.16 Poisson mass distributions for $\mu = 10$ and $\mu = 20$.

small values of μ. However, as the two plots in Figure 5.16 show, the Poisson does resemble the normal distribution for large values of μ. The first plot shows the probability mass function for the Poisson with a mean of 10 and the second plot shows the probability mass function for the Poisson distribution with a mean of 20.

As can be seen from these plots, the normal distribution should be a reasonable approximation to the Poisson distribution for values of μ greater than 10. The normal approximation to the Poisson uses the mean and variance from the Poisson distribution for the normal mean and variance.

Example 5.10

We use the preceding pertussis example to demonstrate the normal approximation to the Poisson distribution. In the pertussis example, we wanted to find the probability of 18 or fewer cases of pertussis, given that the mean of the Poisson distribution was 35.31. This value, 35.31, will be used for the mean of the normal and its square root, 5.942, for the standard deviation of the normal. Since we are using a continuous distribution to approximate a discrete one, we must use the continuity correction. Therefore, we want to find the probability of values less than 18.5. To do this, we convert 18.5 to a z value by subtracting the mean of 35.31 and dividing by the standard deviation of 5.942. The z value is –2.829. The probability of a Z variable being less than –2.829 or –2.83 is found from Table B4 to be 0.0023, close to the exact value of 0.001 given above.

Conclusion

Three of the more useful probability distributions — the binomial, the Poisson, and the normal — were introduced in this chapter. Examples of their use in describing data were provided. The examples also suggested that the distributions could be used to examine whether or not the data came from a particular population or some other population. This use will be explored in more depth in subsequent chapters on interval estimation and hypothesis testing.

EXERCISES

5.1 According to data from NHANES II (NCHS 1992), 26.8 percent of persons 20–74 years of age had high serum cholesterol values (greater than or equal to 240 mg/dL).

a. In a sample of 20 persons ages 20–74, what is the probability that 8 or more persons had high serum cholesterol? Use Table B2 to approximate this value first and then provide a more accurate answer.

b. How many persons out of the 20 would be required to have high cholesterol before you would think that the population from which your sample was drawn differs from the U.S. population of persons ages 20–74?

c. In a sample of 200 persons ages 20–74, what is the probability that 80 or more persons had high serum cholesterol?

5.2 Based on reports from state health departments, there were 10.33 cases of tuberculosis per 100,000 population in the United States in 1990 (NCHS 1992). What is the probability of a health department, in a county of 50,000, observing 10 or more cases in 1990 if the U.S. rate held in the county? What is the probability of fewer than 3 cases if the U.S. rate held in the county?

5.3 Assume that systolic blood pressure for 5-year-old boys is normally distributed with a mean of 94 mmHg and a standard deviation of 11 mmHg. What is the probability of a 5-year-old boy having a blood pressure less than 70 mmHg? What is the probability that the blood pressure of a 5-year-old boy will be between 80 and 100 mmHg?

5.4 Less than 10 percet of the U.S. population is hospitalized in a typical year. However, the per capita hospital expenditure in the United States is generally large — for example, in 1990, it was approximately $975. Do you think that the expenditure for hospital care (at the person level) follows a normal distribution? Explain your answer.

5.5 In Harris County, Texas, in 1986, there were 173 cases of Hepatitis A in a population of 2,942,550 (HCHD 1990). The corresponding rate for the United States was 10.0 per 100,000 population. What is the probability of a rate as low as or lower than the Harris County rate if the U.S. rate held in Harris County?

5.6 Approximately 6.5 percent of women ages 30–49 were iron deficient based on data from NHANES II (LSRO 1989). In a sample of 30 women ages 30–49, 6 were found to be iron deficient. Is this result so extreme that you would want to investigate why the percentage is so high?

5.7 Based on data from the Hispanic Health and Nutrition Examination Survey (HHANES) (LSRO 1989), the mean serum cholesterol for Mexican-American males ages 20 to 74 was 203 mg/dL. The standard deviation was approximately 44 mg/dL. Assume that serum cholesterol follows a normal distribution. What is the probability that a Mexican-American male in the 20–74 age range has a serum cholesterol value greater than 240 mg/dL?

5.8 In 1988, 71% of 15- to 44-year-old U.S. women who have ever been married have used some form of contraception (NCHS 1992). What is the probability that, in a sample of 200 women in these childbearing years, fewer than 120 of them have used some form of contraception?

5.9 In ecology, the frequency distribution of the number of plants of a particular species in a square area is of interest. Skellam (1952) presented data on the number of plants of Plantago major present in squares of 100 square centimeters laid down in grassland. There were 400 squares and the numbers of plants in the squares are as follows:

Plants per Square	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Frequency	235	81	43	18	9	6	4	4

Create a Poissonness plot to examine whether or not these data follow the Poisson distribution.

5.10 The Bruce treadmill test is used to assess exercise capacity in children and adults. Cumming, Everatt, and Hastman (1978) studied the distribution of the Bruce treadmill test endurance times in normal children. The mean endurance time for a sample of 36 girls 4–5 years old was 9.5 minutes with a standard deviation of 1.86 minutes. If we assume that these are the true population mean and standard deviation, and if we also assume that the endurance times follow a normal distribution, what is the probability of observing a 4-year-old girl with an endurance time of less than 7 minutes? The 36 values shown here are based on summary statistics from the research by Cumming et al. Do you believe that these data are normally distributed? Explain your answer.

Hypothetical Endurance Times in Minutes for 36 Girls 4 to 5 Years of Age											
5.3	6.5	7.0	7.2	7.5	8.0	8.0	8.0	8.0	8.2	8.5	8.5
8.8	8.8	8.9	9.0	9.0	9.0	9.0	9.5	9.8	9.8	10.0	10.0
10.6	10.8	11.0	11.2	11.2	11.3	11.5	11.5	12.2	12.4	12.7	13.3

5.11 Seventy-nine firefighters were exposed to burning polyvinyl chloride (PVC) in a warehouse fire in Plainfield, New Jersey, on March 20, 1985. A study was conducted in an attempt to determine whether or not there were short- and long-term respiratory effects of the PVC (Markowitz 1989). At the long-term follow-up visit at 22 months after the exposure, 64 firefighters who had been exposed during the fire and 22 firefighters who were not exposed reported on the presence of various respiratory conditions. Eleven of the PVC exposed firefighters had moderate to severe shortness of breath compared to only 1 of the nonexposed firefighters.

What is the probability of finding 11 or more of the 64 exposed firefighters reporting moderate to severe shortness of breath if the rate of moderate to severe shortness of breath is 1 case per 22 persons? What are two possible confounding variables in this study that could affect the interpretation of the results?

REFERENCES

Blyth, C. R., and H. A. Still. "Binomial Confidence Intervals." *Journal of the American Statistical Association* 78:108–116, 1983.

Boyer, C. B. *A History of Mathematics*. Princeton University Press, 1985, p. 569.

Cumming, G. R., D. Everatt, and L. Hastman. "Bruce Treadmill Test in Children: Normal Values in a Clinic Population." *The American Journal of CARDIOLOGY* 41:69–75, 1978.

Harris County Health Department (HCHD), Mark Canfield, editor. *The Health Status of Harris County Residents: Births, Deaths and Selected Measures of Public Health, 1980–86*, 1990.

Hoaglin, D. C. "A Poissonness Plot." *The American Statistician* 34:146–149, 1980.

Life Sciences Research Office (LSRO), Federation of American Societies for Experimental Biology: *Nutrition Monitoring in the United States — An Update Report on Nutrition Monitoring*. Prepared for the U.S. Department of Agriculture and the U.S. Department of Health and Human Services. DHHS Pub. No. (PHS) 89–1255, 1989.

Markowitz, J. S. "Self-Reported Short- and Long-Term Respiratory Effects among PVC-Exposed Firefighters." *Archives of Environmental Health* 44:30–33, 1989.

McPherson, R. S., M. Z. Nichaman, H. W. Kohl, D. B. Reed, and D. R. Labarthe. "Intake and food sources of dietary fat among schoolchildren in The Woodlands, Texas." *Pediatrics* **88**(4): 520–526, 1990.

National Center for Health Statistics. *Health, United States, 1991 and Prevention Profile*. Hyattsville, MD: Public Health Service, DHHS Pub. No. 92–1232, 1992, Tables 50 and 70.

Public Citizen Health Research Group. "Unnecessary Cesarean Sections: Halting a National Epidemic." *Public Citizen Health Research Group Health Letter* 8(6):1–6, 1992.

Samuels, M. L., and T. C. Lu. "Sample Size Requirements for the Back-of-the-Envelope Binomial Confidence Interval." *The American Statistician* 46:228–231, 1992.

Schader, M., and F. Schmid. "Two Rules of Thumb for the Approximation of the Binomial Distribution by the Normal Distribution." *The American Statistician* 43:23–24, 1989.

Skellam, J. G. "Studies in Statistical Ecology." *Biometrika* 39:346–362, 1952.

Stigler, S. M. *American Contributions to Mathematical Statistics in the Nineteenth Century*. New York: Arno Press, 1980.

Student. "On the Error of Counting with a Haemacytometer." *Biometrika* 5:351–360, 1907.