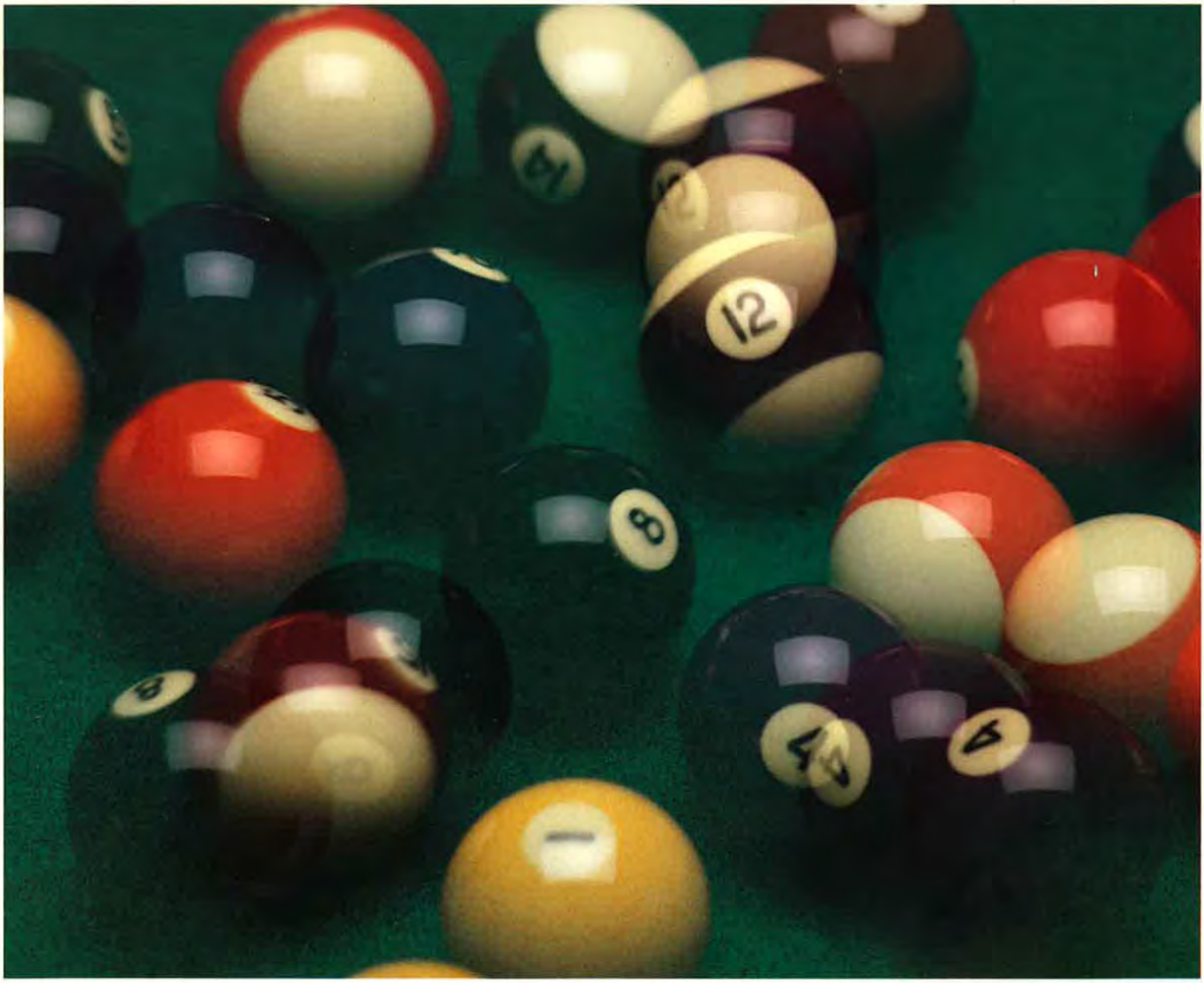
ADVANCED ALGEBRA

Probability Models

P. HOPFENSPERGER, H. KRANENDONK, R. SCHEAFFER

DATA-DRIVEN MATHEMATICS



DALE SEYMOUR PUBLICATIONS®

Probability Models

DATA-DRIVEN MATHEMATICS

Patrick Hopfensperger, Henry Kranendonk, and Richard L. Scheaffer

Dale Seymour Publications®
White Plains, New York

This material was produced as a part of the American Statistical Association's Project "A Data-Driven Curriculum Strand for High School" with funding through the National Science Foundation, Grant #MDR-9054648. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Managing Editor: Alan MacDonell

Senior Math Editor: Carol Zacny

Project Editor: Nancy R. Anderson

Production/Manufacturing Director: Janet Yearian

Production Coordinator: Roxanne Knoll

Design Manager: Jeff Kelly

Text and Cover Design: Christy Butterfield

Cover photo: Stephen Frisch

This book is published by Dale Seymour Publications®, an imprint of Addison Wesley Longman, Inc.

Dale Seymour Publications
10 Bank Street
White Plains, New York 10602
Customer Service: 800-872-1100

Printed in the United States of America.

Order number DS21179

ISBN 1-57232-240-3

1 2 3 4 5 6 7 8 9 10-ML-03 02 01 00 99 98

This Book Is Printed On Recycled Paper

Authors

Patrick Hopfensperger
Homestead High School
Mequon, Wisconsin

Henry Kranendonk
Rufus King High School
Milwaukee, Wisconsin

Richard Scheaffer
University of Florida
Gainesville, Florida

Consultants

Jack Burrill
National Center for Mathematics Sciences Education
University of Wisconsin-Madison
Madison, Wisconsin

Emily Errthum
Homestead High School
Mequon, Wisconsin

Maria Mastromatteo
Brown Middle School
Ravenna, Ohio

Vince O'Connor
Milwaukee Public Schools
Milwaukee, Wisconsin

Jeffrey Witmer
Oberlin College
Oberlin, Ohio

Data-Driven Mathematics Leadership Team

Miriam Clifford
Nicolet High School
Glendale, Wisconsin

Kenneth Sherrick
Berlin High School
Berlin, Connecticut

Richard Scheaffer
University of Florida
Gainesville, Florida

James M. Landwehr
Bell Laboratories
Lucent Technologies
Murray Hill, New Jersey

Gail F. Burrill
National Center for Mathematics Sciences Education
University of Wisconsin-Madison
Madison, Wisconsin

Acknowledgments

The authors thank the following people for their assistance during the preparation of this module:

- The many teachers who reviewed drafts and participated in field tests of the manuscripts
- The members of the *Data-Driven Mathematics* leadership team, the consultants, and the writers
- Kathryn Rowe and Wayne Jones for their help in organizing the field-test process and the Leadership Workshops

Table of Contents

About *Data-Driven Mathematics* vi

Using This Module vii

Unit I: Random Variables and Their Expected Values

Lesson 1: **Probability and Random Variables** 3

Lesson 2: **The Mean as an Expected Value** 10

Lesson 3: **Expected Value of a Function of a Random Variable** 16

Lesson 4: **The Standard Deviation as an Expected Value** 21

Assessment: **Lessons 1–4** 28

Unit II: Sampling Distributions of Means and Proportions

Lesson 5: **The Distribution of a Sample Mean** 35

Lesson 6: **The Normal Distribution** 44

Lesson 7: **The Distribution of a Sample Proportion** 53

Assessment: **Lessons 5–7** 61

Unit III: Two Useful Distributions

Lesson 8: **The Binomial Distribution** 67

Lesson 9: **The Geometric Distribution** 76

Assessment: **Lessons 8 and 9** 84

About *Data-Driven Mathematics*

Historically, the purposes of secondary-school mathematics have been to provide students with opportunities to acquire the mathematical knowledge needed for daily life and effective citizenship, to prepare students for the workforce, and to prepare students for postsecondary education. In order to accomplish these purposes today, students must be able to analyze, interpret, and communicate information from data.

Data-Driven Mathematics is a series of modules meant to complement a mathematics curriculum in the process of reform. The modules offer materials that integrate data analysis with high-school mathematics courses. Using these materials will help teachers motivate, develop, and reinforce concepts taught in current texts. The materials incorporate major concepts from data analysis to provide realistic situations for the development of mathematical knowledge and realistic opportunities for practice. The extensive use of real data provides opportunities for students to engage in meaningful mathematics. The use of real-world examples increases student motivation and provides opportunities to apply the mathematics taught in secondary school.

The project, funded by the National Science Foundation, included writing and field testing the modules, and holding conferences for teachers to introduce them to the materials and to seek their input on the form and direction of the modules. The modules are the result of a collaboration between statisticians and teachers who have agreed on statistical concepts most important for students to know and the relationship of these concepts to the secondary mathematics curriculum.

Using This Module

Why the Content Is Important

Data analysis is concerned with studying the results of an investigation that has already taken place, with the hope of discovering some patterns in the data that might lead to new insights into the behavior of one or more variables. Probability is concerned with anticipating the future, with the hope of discovering models that might allow the prediction of outcomes not yet seen. Of course, we cannot predict with certainty and so possible outcomes are generally stated along with their chances of occurring.

There is a connection between data and probability since the probabilities used for anticipating future events often come from the analysis of past events. Thus, a survey that says 60% of drivers do not wear seat belts serves as the basis for calculating the probability distribution for the number of drivers, out of the next ten observed, who are not wearing seat belts.

There is also a connection between the key components of describing distributions of data and the key components of describing probability distributions. The mean of the data parallels the expected value of the probability distribution, but notice the change in language from something we see as a fact to something we merely anticipate. Variation in data and in probability distributions is often measured by the standard deviation, but the calculation becomes the expected value of a function of a random variable in the latter case. Shapes of data distributions and probability distributions are described by the same terms–symmetric and skewed–so the context of such descriptions must be made clear.

In this module, students will learn about the connections between data analysis and probability. The emphasis is on the development of basic concepts of probability distributions, as contrasted with probability from counting rules, and the use of standard models for these distributions. The value of having such standard models is that you need study only a few probability distributions in order to solve a wide variety of probability problems. Students will see that a lot of mileage is obtained from the normal, binomial, and geometric models.

The skills required for working through this module are mainly those of beginning algebra, except that an infinite series is introduced in Lesson 9. Experience with simulation would be helpful, as many ideas are introduced with this approach.

Random Variables and Their Expected Values

LESSON 1

Probability and Random Variables

How many children are in a typical American family?

What is the probability of randomly choosing a family with two children?

What is a random variable?

OBJECTIVES

Understand the relative-frequency concept of probability.

Define random variables.

According to the U.S. Bureau of the Census, the number of children under 18 years of age per family has a distribution as given on the table below. A "family" is defined as a group of two or more persons related by birth, marriage, or adoption, residing together in a household. In which category does your family belong?

INVESTIGATE

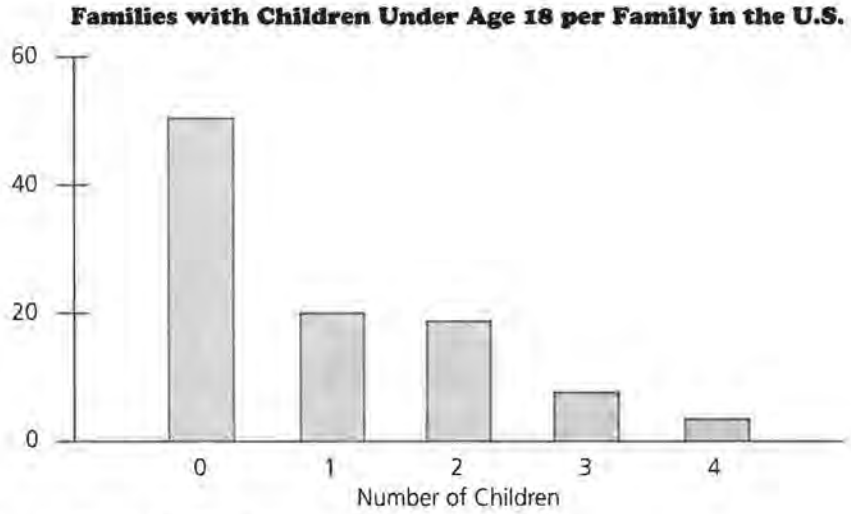
Family Size

In reality, some families have more than four children under the age of 18. However, the number of such families is very small and their percent would be very small compared to the percents in this table. Thus, we can describe the essential features of the number of children per family by using this simplified table as a *model* of reality.

Number of Children	Percent of Families
0	51
1	20
2	19
3	7
4	3

Discussion and Practice

1. The first line of data in the table above is interpreted to mean that 51% of the families in the United States have no children under the age of 18.
   a. What percent of families have one child under the age of 18?
   b. What percent of families have at least one child under the age of 18?
   c. How would you interpret the 7% for the "3 children" category?
   d. What is the sum of the percents in the table? Explain why this is an appropriate value for the sum.
2. A bar graph of the data on the number of children per family is shown below.
   a. What does the height of the bar over the 2 represent?
   b. What percent of families have at most two children under the age of 18?
   c. Describe in words the distribution of children per family.



Random Variables

The discussion above makes use of the data table to describe one aspect of families in the United States. Suppose A. C. Nielsen, the company that provides ratings of TV shows, is planning to select a random sample of families from across the country. In that case, these same percents can be used as probabilities so that Nielsen can anticipate how many children under the age of 18 they might encounter in the sample.

3. Suppose Nielsen is to select one family at random. What is the approximate probability that the selected family will have

   a. exactly one child under the age of 18?

   b. at least one child under the age of 18?

   c. at most two children under the age of 18?

   d. either two or three children under the age of 18?

   e. exactly five children under the age of 18?

In your past work, symbols have helped you to communicate mathematical statements more clearly and more concisely. Symbols can also help to clarify probability statements. In the situation above, the numerical outcome of interest is "the number of children under the age of 18 in a randomly selected U.S. family." Instead of writing this long statement each time we need it, why not just call it C? Then, C = the number of children under the age of 18 in a randomly selected U.S. family. From the data table, you can see that the probability that C is equal to 1 is 0.20, or 20%. It is cumbersome to write this probability statement in words, so we use a shorthand notation for the statement. The symbolic statement is $P(C = 1) = 0.20$.

When probability statements involve intervals of values for C, the symbolic form makes use of inequalities. For example, the probability that a randomly selected family has "at most one child under the age of 18" implies that the family has "either 0 or 1 child under the age of 18." This can be written as

$$P(C = 0) + P(C = 1) = P(0 \leq C \leq 1) = 0.51 + 0.20 = 0.71.$$

4. Write a symbolic statement for each of the statements in Problem 3.

5. Use the data table on page 3 to find the following probabilities.

   a. $P(C = 0)$

   b. $P(C \leq 2) = P[(C = 0) \text{ or } (C = 1) \text{ or } (C = 2)]$

   c. $P(C \geq 3)$

   d. $P(1 \leq C \leq 3)$

6. Write each of the following symbolic statements in words.

   a. $P[(C = 1) \text{ or } (C = 3)]$

   b. $P(C \geq 2)$

c. $P(2 \le C \le 4)$

d. $P[(C \le 2) \text{ or } (C \ge 4)]$

7. The ***complement*** of an event includes all possible outcomes except the ones in that event. For each of the symbolic statements in Problem 6, write a symbolic statement for the *complement of the event* in question. Find the probability of each complement.

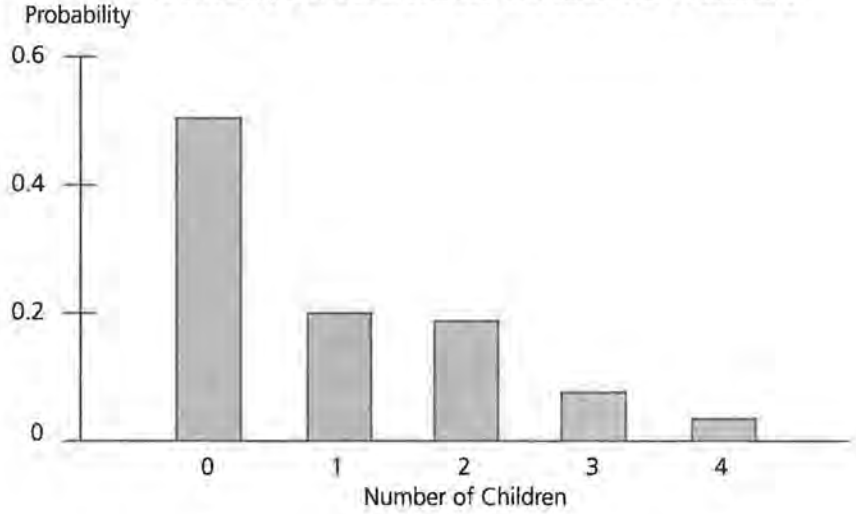
8. Write "the probability that there are no more than three children in a randomly selected family" in symbolic form and find a numerical answer for this probability.

The symbols like C used to represent numerical outcomes from chance processes are called ***random variables***. Random variables are the basic building blocks for working with probability in scientific investigations. Probability ***distributions*** for random variables can be conveniently displayed in a two-column table like the one shown below for the random variable C, the "number of children under 18 in a randomly selected family."

C	$P(C)$
0	0.51
1	0.20
2	0.19
3	0.07
4	0.03

The probability distribution for a random variable can also be displayed in a bar graph, like the one shown below for the random variable C.

Probability Distribution for the Random Variable *C*



9. Study the relationship between the probability distribution as expressed in the table and as expressed in the graph.

a. Describe in words the shape of the probability distribution shown above.

b. What are the differences between the graphs in Problem 2 and Problem 8? Describe the different purposes they serve.

c. Add the column of probabilities in the table for the random variable *C*. What should be the sum of the probabilities in a complete probability distribution? Explain why this must be the case.

Practice and Applications

10. Consider another relative frequency distribution that can be turned into a probability distribution for a random variable. According to the *Statistical Abstract of the United States* (1996), the number of motor vehicles available to American households is given by the percents shown in the following table. A "household" is defined as all persons occupying a housing unit such as a house, an apartment, or a group of rooms. Notice the difference between a family and a household.

Number of Motor Vehicles per Household	Percent of Households
0	1.4
1	22.8
2	43.7
3	21.5
4	10.6

Note: Very few households have more than four motor vehicles.

The data in the table considers any transportation device that requires a motor-vehicle registration by the state in which it is located. For convenience, however, we will refer to these motor vehicles as "cars."

a. The sum of the percents in the table is 100. Does that mean that no household has more than four cars? Explain.

b. Define a random variable *Y* to be the number of cars available to a randomly selected American household. Construct the probability distribution for *Y* in table form.

c. Construct a bar graph that represents the probability distribution for Y. Describe the shape of this distribution.

11. An automobile manufacturer is planning to conduct a survey on what Americans think about automobile repairs. What is the chance that a randomly selected household in the poll has

a. no cars?

b. exactly one car?

c. at least one car?

d. at most two cars?

e. exactly three cars?

12. Write each of the statements in Problem 11 in symbolic form.

13. With Y defined as in Problem 10, find each of the following probabilities and write the symbolic statement in words.

a. $P(Y = 2)$

b. $P(Y \geq 0)$

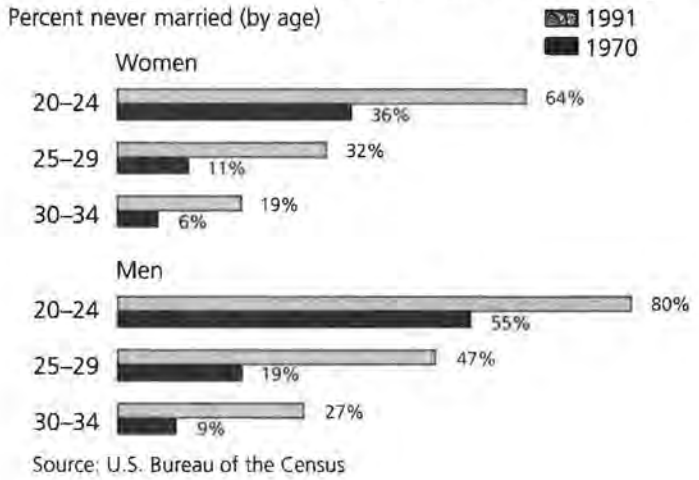
c. $P(Y \leq 3)$

d. $P(1 \leq Y \leq 3)$

14. For the first randomly selected household contacted, what is the probability that the household has at least one car? Write a symbolic statement for this probability.

15. The graph below provides information on how young adults are postponing marriage.

Increasing Numbers of Young Adults Are Delaying Marriage



Source: U.S. Bureau of the Census

a. Do men tend to postpone marriage longer than women do? Use the data from the graph to support your answer.

b. Suppose a 1991 survey randomly sampled women between the ages of 20 and 24. What is the probability that the first such woman sampled was married?

c. Suppose a 1991 survey randomly sampled men between the ages of 25 and 29. What is the probability that the first such man sampled had never married?

d. From these data, can we answer the following question? Explain. "What is the probability that an adult randomly selected in a 1991 survey was under the age of 34 and had never married at the time of his or her selection?"

SUMMARY

A display, such as a table or a graph, showing the numerical values that a variable can take on and the percent of time that the variable takes on each value is called the *distribution* of that variable. If possible values of the variable are randomly selected, the variable is called a *random variable* and the percents attached to the numerical values give the probability distribution for that variable.

LESSON 2

The Mean as an Expected Value

What is the average number of children per family in America?

In a randomly chosen family, how many children would you expect to see?

How does the mean number of children per family compare to the mean family size?

OBJECTIVE

Understand how to compute and interpret the mean of a probability distribution.

An average, such as the arithmetic *mean* or simply the mean, is a common measure of the center of a set of data. The mean score of your quizzes in mathematics is, no doubt, an important part of your grade in the course. The mean age of residences in your neighborhood helps insurance companies figure out how much to charge for fire insurance. The mean amount paid by families for typical goods and services this year as compared to last year determines the rate of inflation. In this lesson, we will look at means of distributions of data to discover how they relate to means of probability distributions.

INVESTIGATE

How would you calculate the mean score of your quizzes in mathematics? In what other situations might you need to calculate the mean for a set of data?

Discussion and Practice

1. A football team played nine games this season, scoring 12 points in each of three games and 21 points in each of the other six games.

a. Construct a bar graph for the points scored, with the values for the variable "points scored" on the horizontal axis.

b. What is the mean number of points scored per game for this team? Explain how you found this mean.

c. Mark the value of the mean on the horizontal axis of the bar graph. Is the mean closer to 12 or to 21?

2. Suppose you knew that the team scored 12 points in $\frac{1}{3}$ of its games and 21 points in $\frac{2}{3}$ of its games, but you were not told how many games the team played.

a. Construct a bar graph for these data. How does it compare to the one in Problem 1a?

b. Can you still calculate the mean number of points per game? If so, what is it? Discuss how you arrived at this answer.

c. Mark the mean on the horizontal axis of the bar graph.

3. The team is expected to perform next year about as well as it performed this year. That is, the probability of scoring 12 points in a game is about $\frac{1}{3}$, while the probability of scoring 21 points in a game is about $\frac{2}{3}$. For a randomly selected game, how many points would you expect the team to score?

4. A mean computed from a probability distribution—an anticipated distribution of outcomes—is called an ***expected value***. Discuss why you think this terminology is used. Does the terminology seem appropriate?

5. Instead of a randomly selected game from next year's schedule, suppose we consider the game against the best team in the league. Would that change your opinion on the team's expected number of points scored? Why or why not?

Expected Value

Recall that one of the first numerical summaries of a set of data that you studied was the mean, used as a measure of center. We now review the calculation of the mean by working through an example. A survey of a class of 20 students reveals that 4 have no pets, 10 have one pet, and 6 have two pets. The data are shown in the table below.

Number of Pets	Number of Students	Total Number of Pets
0	4	0
1	10	10
2	6	12

The mean number of pets per student can be calculated in a variety of ways.

6. What is the mean number of pets per student? Discuss your calculation method with someone else in the class who used a different method.

7. Suppose we do not know how many students were surveyed, but we do know the percent of students who had each number of pets.

a. What percent of the students have no pets? One pet? Two pets? Add a column to the table for these percents.

b. Based on the percents in part a, find the mean number of pets per student surveyed. Explain how you arrived at your answer.

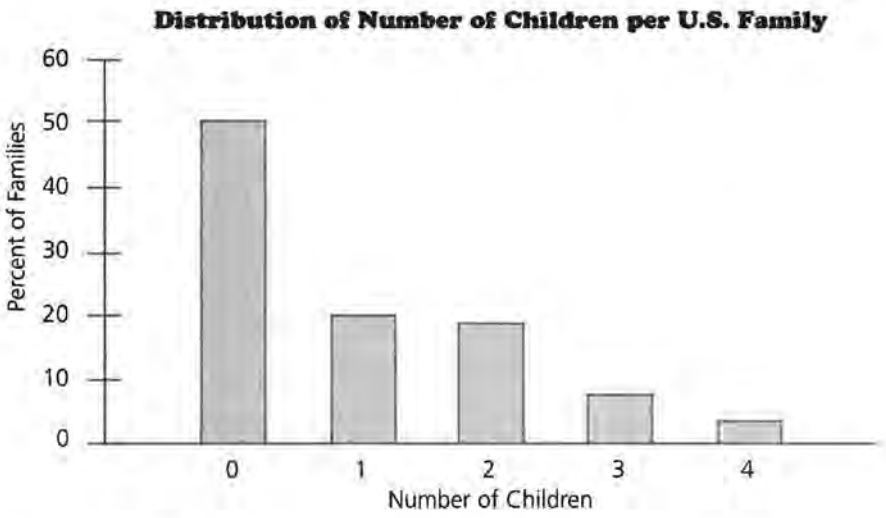
c. How does the answer compare to your answer for Problem 6? Should the answers be the same?

8. We now return to the Lesson 1 data on the number of children under the age of 18 per U.S. family.

Number of Children	Percent of Families
0	51
1	20
2	19
3	7
4	3

a. Use what you just learned to calculate the mean number of children per family in the U.S.

b. Find the mean on the horizontal scale of the bar graph for these data provided in the following graph. Is the mean in the center of the distribution? Why or why not?



9. The A. C. Nielsen Company randomly selects families for use in estimating the ratings of TV shows. For each randomly selected family, how many children would we expect to see? That is, what is the expected value of C, the number of children in a randomly selected family? Show how you found your answer.

10. Explain why the terminology changed from *mean* number of children per family in Problem 8 to *expected* number of children per family in Problem 9.

11. The calculation of an expected value often results in a decimal. That is, the answer is not always an integer.

- **a.** Explain why the decimal part of the expected number of children per family makes sense as an expected value, even though we cannot see a fraction of a child in any one family.
- **b.** How many children in all would we expect to see in a random sample of 100 families?
- **c.** How many children would we expect to see in a random sample of 2500 families?
- **d.** If Nielsen really expects opinions from about 4000 children under the age of 18, how many families should be in the sample?

You now have the tools to develop a general expression for the expected value of a random variable.

12. Suppose a random variable X can take on values $x_1, x_2, \ldots, x_k$ with respective probabilities $p_1, p_2, \ldots, p_k$. That is, $P(X = x_i) = p_i$ for values of i ranging from 1 to k.

a. Write a symbolic expression for the expected value of X. Explain the reasoning behind this expression.

b. A commonly used symbol for the expected value of X is $E(X)$, and $E(X)$ is expressed as a sum. The Σ symbol tells you to add the terms that follow the symbol, starting with the term indicated by the integer below the Σ and ending with the term indicated by the integer above the Σ. Thus,

$$\sum_{i=1}^{3} x_i p_i = x_1 p_1 + x_2 p_2 + x_3 p_3$$

Replace the question marks in the expression below with numerical values to indicate the range of summation.

$$E(X) = \sum_{i=?}^{?} x_i (?)$$

Practice and Applications

The following table shows the distribution of household sizes for U.S. households.

Number of Persons per Household	Percent of Households
1	25
2	32
3	17
4	16
5	7
6	2
7	1

Note: Households of more than 7 persons are very rare.

13. Suppose Nielsen is randomly sampling households in order to produce TV ratings. Let Y denote the size of a randomly selected household.

a. Find the expected value of Y. Compare it to the expected value of C found in Problem 9.

b. If Nielsen randomly selects 1000 households, how many people would these households be expected to contain?

c. If Nielsen really expects 4000 people in the survey, how many households should be sampled?

14. Construct a bar graph of the probability distribution for Y. Mark the expected value of Y on the bar graph. Is the expected value in the center of the possible values for Y? Why or why not?

SUMMARY

The distribution of a variable is determined by the numerical values that the variable can take on, along with the proportion of times that each numerical value occurs. The ***mean*** of the distribution can be calculated from this information. If the proportions of times that the numerical values occur are interpreted as probabilities, then the mean is called the ***expected value*** of a probability distribution. The mean of a data set describes something that has already happened, while an expected value anticipates what might happen in the future.

LESSON 3

Expected Value of a Function of a Random Variable

How much would you expect to pay to feed a pet for a week?

What is a "fair" game?

How do business decisions depend on the concept of a fair game?

OBJECTIVES

Find expected values of certain functions of random variables.

Understand fair games.

As you have seen in previous work in mathematics and science, it is often convenient to express one variable as a function of another. Most goals in basketball are worth 2 points; hence, the number of points a player scores in a game, excluding free throws and 3-point goals, can be written as a function of the number of goals made.

Charlotte takes about 20 shots from inside the 2-point area during a game and expects to make about 60% of them. How many points can she expect to get from these goals?

INVESTIGATE

Expected Value of a Function

In practical applications of probability, the random variables are often written as functions of other random variables. Suppose the table below gives estimates of the probabilities of a randomly selected student having 0, 1, or 2 pets.

Number of Pets	Probability
0	0.2
1	0.5
2	0.3

If X represents the number of pets per randomly selected student, then $E(X)$ can be calculated from the information in the table.

Discussion and Practice

1. Calculate $E(X)$ for the pet example.
2. Suppose that it costs around \$20 per week to feed a pet. We now want to study the probability distribution for a new random variable Y, the cost of pet feeding per week.
   a. Write the probability distribution for Y in table form, based on the information provided above on number of pets per student.
   b. Use the results of Problem 2a to calculate $E(Y)$, the expected weekly pet feeding cost per student.
3. Another way to find the expected value of Y is to observe that $Y = 20X$.
   a. Use the formula for $E(Y)$ and $E(X)$ to show that $E(Y) = 20E(X)$.
   b. Calculate $E(Y)$ by using the result in Problem 3a. Compare this result with your answer to Problem 2b.
4. Sam has a job mowing lawns. He expects to mow 8 lawns per week. He charges \$15 per lawn, but it costs him about \$20 per week to keep the lawn mower in good repair and full of fuel. What is Sam's expected profit per week?
5. Show that, in general, if $Y = aX + b$ then $E(Y) = aE(X) + b$.
   a. Begin by writing the formula for $E(Y)$ as a summation, assuming Y can take on values $y_1, y_2, \ldots, y_k$, with respective probabilities $p_1, p_2, \ldots, p_k$.
   b. Substitute $y_i = ax_i + b$ inside the summation.
   c. Use the Distributive Property to write the terms inside the summation as a sum of two terms.
   d. Use the properties of sums to write the summation as a term involving $E(X)$.

Practice and Applications

6. Refer to Problem 13 in Lesson 2. Suppose the cost to the Nielsen Company for connecting a family to their system is a flat rate of \$500 per household plus \$100 for every family member in the household. How much should Nielsen expect to pay per family for connection charges?

7. Refer to Problems 8 and 9 in Lesson 2. The Gallup organization wants to sample children under the age of 18 and ask them about their attitudes toward school. It cannot sample children directly but it can sample families. It takes about 10 minutes to question the family about the status of their children and about 30 additional minutes for each interview conducted. How much time, on the average, should Gallup allow for each family sampled?

Fair Games

In the town raffle, a drawing is to take place for a radio worth about \$100. Two hundred tickets will be sold for \$1 each. The tickets are mixed in a drum and one ticket is randomly selected for the winning prize. If you buy one ticket, let's analyze what happens to G, the amount you gain.

There are two possible outcomes: you win or you lose. If you lose you have lost \$1, which can be called a gain of –1. If you win, however, you gain \$100 minus the \$1 you paid to play, for a net gain of \$99. So the probability distribution for G is as shown in the table.

G	$P(G)$
–1	199/200
99	1/200

By the rules of expected value, your expected gain is

$$E(G) = -1\left(\frac{199}{200}\right) + 99\left(\frac{1}{200}\right) = -\frac{1}{2}.$$

You can expect to lose a half dollar for every play of such a game. Would you call this a fair game?

8. Write a reasonable definition of a fair game.

9. What would you be willing to pay for a drawing like the one above to make the drawing fair in the sense of expected gain?

10. If the game is fair for you as a player, do the people running the drawing make any money? Do you see a reason why most games are not fair?

11. There is another way to assess the expected gain for the game described above. Suppose we define W as the amount you win. Then, your gain can be written as a function of W.

- **a.** Find the probability distribution of W for the game described at the beginning of this investigation, in which you pay \$1 to play.
- **b.** Find the expected value of W.
- **c.** Write the player's gain G as a function of W, with G and W as defined for the game above.
- **d.** Use $E(W)$ to find $E(G)$. Does the answer agree with what we found earlier? Which method seems easier?

12. N tickets are sold for a drawing that will have one randomly selected winner. The payoff is an amount A. Each ticket sells for an amount C.

- **a.** Find the probability distribution for the winnings of a player who buys one ticket.
- **b.** Find the expected winnings for a player who buys one ticket.
- **c.** How much should each ticket cost if this is to be a fair game?
- **d.** Do the answers to Problems 12b and 12c seem reasonable? Explain.

13. An insurance company insures a car for \$20,000. The one-year premium paid for the insurance is denoted by r. The company has records on drivers and cars of the type insured here and estimates that they will sustain a total loss with probability 0.01 and a 50% loss with probability 0.05. All other partial losses are ignored.

- **a.** Find the probability distribution for the amount the company pays out.
- **b.** Find the company's expected gain if r = \$1000.
- **c.** What should the company charge as a premium to make this a "fair game"? Can the company actually do this? Explain.

SUMMARY

Many practical applications of probability involve finding expected values of functions of random variables. For linear functions of the form $Y = aX + b$ for constants a and b,

$$E(Y) = aE(X) + b.$$

LESSON 4

The Standard Deviation as an Expected Value

Do most households have about the same number of cars, or is there a great deal of variation from household to household?

Is the number of persons per family more variable than the number of children per family?

How can you measure variation in a probability distribution?

OBJECTIVE

Understand how to compute and interpret the standard deviation of data and probability distributions.

In data analysis, once we have a measure of center, it is important to develop a measure of ***variation***, or spread, of the data to either side of the center. One useful measure of variability is the standard deviation, a value you may have encountered in lessons on data analysis. We now develop that same measure of spread for probability distributions.

A ***deviation*** is the distance between an observed data point and the mean of the distribution of data. The average of the squared deviations has a special name, ***variance***. The square root of the variance is called the ***standard deviation***. The standard deviation has important practical uses in probability and statistics, some of which we will see in future lessons of this unit.

INVESTIGATE

Recall the data in Lesson 2 regarding the number of pets students have. The survey of 20 students revealed that 4 have no pets, 10 have one pet, and 6 have 2 pets. A tabular array for

these data follows. How could you calculate the numbers in the third column?

Number of Pets	Number of Students	Percent of Students
0	4	20
1	10	50
2	6	30

Discussion and Practice

The mean number of pets per student is 1.1. Do you remember how we determined this? The standard deviation is a special function of the variable "number of pets" and can be calculated by making use of what we learned in Lessons 2 and 3.

1. Use the following steps to find the variance of the number of pets per student.
   a. Add a column of "deviations from the mean" to the table. What would you say is a "typical" deviation?
   b. Find the average of the deviations from the mean.
   c. Add a column of "squared deviations from the mean" to the table.
   d. Find the average of the squared deviations from the mean, called the *variance*.
2. The *standard variation* is the square root of the variance. Find the standard deviation of the number of pets per student. Is this number close to what you chose as a typical deviation in Problem 1a?
3. Draw a bar graph of the data on the number of pets per student given above.
   a. Mark the mean of this distribution on the graph.
   b. Mark off a distance of one standard deviation above, that is, to the right of, the mean.
   c. Mark off a distance of one standard deviation below, that is, to the left of, the mean.
   d. What fraction of the 20 data values are inside the interval from one standard deviation below the mean to one standard deviation above the mean?
4. Sketch another bar graph, still using data values of 0, 1, and 2 but choosing frequencies which would have greater

standard deviation than the one in Problem 3. Explain what feature of the bar graphs is measured by standard deviation.

Consider again the distribution of the number of children under the age of 18 in U.S. families as given in the table below.

Number of Children	Percent of Families
0	51
1	20
2	19
3	7
4	3

5. We now study this distribution using what you learned earlier in this lesson.

a. Calculate the standard deviation of the number of children per family. Recall that the mean was 0.91.

b. Sketch a bar graph of this distribution. Mark the mean number of children per family on the graph.

c. Mark off a distance of one standard deviation to both sides of the mean.

d. What percent of families would have a number of children inside the interval marked off in Problem 5c? How does this value compare with the answer to Problem 3d?

6. Sometimes the standard deviation is referred to as a "typical" deviation between a data point and the mean. Is this a fitting description? Explain.

7. The A. C. Nielsen Company plans to randomly select a large number of families to be used in collecting data for rating TV shows. Let C represent the random variable "number of children under the age of 18 in a randomly selected U.S. family."

a. Find the standard deviation we would expect for C, based on the available data and the fact that the expected value is 0.91.

b. Is there any difference between the numerical values for standard deviations calculated in Problems 5 and 7a?

c. Is there any difference in interpretation between the standard deviations calculated in Problems 5 and 7a? Explain.

You can now make the transition from working with numbers to working with symbols. The goal is to develop a formula for the standard deviation as calculated from a probability distribution.

8. Suppose a random variable X can take on the values $x_1, x_2, \ldots, x_n$ with respective probabilities $p_1, p_2, \ldots, p_n$. Write a symbolic expression for the standard deviation of X as an expected value.

Practice and Applications

9. Data on automobiles per family in the U.S. are given below.

Number of Cars	Percent of Households
0	1.4
1	22.8
2	43.7
3	21.5
4	10.6

a. Calculate the expected value and standard deviation of the number of cars per household that would be expected in a random sample of households from the U.S.

b. What percent of the households have a number of cars within one standard deviation of the mean?

c. Suppose you are allowed to use only the mean and standard deviation to describe these data in a newspaper article to be read by people who are not familiar with these terms. Write such a description.

10. The distribution of the number of persons per household in the U.S. is given in the following table.

Number of Persons per Household	Percent of Households
1	25
2	32
3	17
4	16
5	7
6	2
7	1

a. Sketch a bar graph for the distribution of the number of persons per household. Compare this distribution with the distribution of the number of children per family. Which will have the greater standard deviation? Explain why without calculating the standard deviation for the number of persons per household.

b. Calculate the standard deviation of the number of persons per household. Does it confirm your answer to Problem 10a?

11. In the tables showing the number of children per family and the number of people per household, the greatest value shown in the tables is not the greatest possible value. That is, there can be more than 4 children in a family and there can be more than 7 people in a household. If more accurate data on large families were available, what effect would that have on the calculated values of the standard deviations? Explain.

12. Looking at all the distributions seen so far in this lesson, for which does the standard deviation seem to be the best as a measure of a typical deviation from the mean? For which does it seem to be the worst? Explain.

13. The table below shows the percents of sports shoes of different types that are sold to various age groups.

Age of User	Gym Shoes	Jogging Shoes	Walking Shoes
Under 14	39.3	8.8	3.3
14 to 17	10.7	11.7	1.9
18 to 24	8.5	8.4	2.7
25 to 34	13.2	22.3	12.2
35 to 44	11.4	24.1	16.2
45 to 64	11.6	19.5	36.6
65 and over	5.3	5.2	27.1

Source: *Statistical Abstract of the United States*, 1993–94

a. Construct meaningful plots of the three age distributions. Comment on their differences. Which has the greatest mean? Which has the greatest standard deviation? You might begin by choosing a meaningful age to represent each of the shoe categories. Then, the data will look more like what we have been studying in this lesson and can be plotted as a bar graph.

b. Approximate the median age of user for each of the three shoe types.

c. It is difficult to calculate the mean age of each user since the ages are given in intervals. For each age group, select a single value which you think best approximates the ages in that interval. Using those selected values, approximate the mean age of user for each of the three shoe types. How do the mean ages compare with the median ages?

d. Using the ages per interval selected in Problem 13c, approximate the standard deviation of ages for each of the three shoe types.

e. Would manufacturers of sports shoes find these means and standard deviations to be useful summaries of the age distributions? Write a summary of these age distributions for a publication on shoe sales, assuming the audience knows very little about statistics.

14. This lesson began with a discussion of the distribution of the number of pets found in a sample of students. In Lesson 3, we assumed that it cost \$20 per week to feed each pet. The distribution of Y, the weekly cost of feeding pets, is shown in this table.

Y	$P(Y)$
0	0.2
20	0.5
40	0.3

a. Use this distribution to find the standard deviation of Y. You may use the fact that the expected value of Y is \$22.

b. Compare the standard deviation of Y to the standard deviation of the number of pets per student, found in Problem 2 to be 0.70. Do you see a simple rule for relating the standard deviation of Y to that of X, the number of pets per student?

15. Suppose a random variable, X, has standard deviation denoted by σ, the Greek letter s. A new random variable is constructed as $Y = aX + b$.

a. What is the standard deviation of Y in terms of σ? Show why this is true by making use of the formula for standard deviation.

b. Suppose the number of persons per household has a mean of 2.6 and a standard deviation of 1.4. Each mem-

ber of a sampled household is to be interviewed by a pollster at a cost of $30 per interview. What are the expected value and standard deviation of the cost of interviewing a randomly selected household? Would this cost exceed $100 very often?

SUMMARY

For distributions of data and probability distributions of random variables the center is often measured by the mean or expected value and the spread by the ***standard deviation***. The standard deviation measures a "typical" deviation between a possible data point and the mean. Most of the data points usually lie within one standard deviation of the mean.

For probability distributions, the standard deviation can be written as an expected value of a function of the underlying random variable. This measure will be used extensively in future lessons of this module.

Lessons 1–4

1. The following table shows the distribution of family sizes for U.S. families.

Number of Persons per Household	Percent of Households
1	25
2	32
3	17
4	16
5	7
6	2
7	1

Note: Households of more than 7 persons are very rare.

Describe the distribution of household size and compare it to the distribution of number of children per family.

2. Provide a reasonable definition for a random variable whose distribution can be approximated from these data on number of persons per household.
3. For a randomly selected family from the U.S., find the probability that the number of persons in the household is
   a. 2 or more.
   b. more than 2.
   c. at least 3.
   d. no more than 3.
   e. between 2 and 4, inclusive.
   f. more than 1, but less than 6.
4. Write each of the statements in Problem 3 in symbolic form.
5. According to the U.S. Bureau of the Census, the number of cars available to American households is given by the following percents.

Number of Cars per Household	Percent of Households
0	1.4
1	22.8
2	43.7
3	21.5
4	10.6

A random sample of households is to be selected to participate in a study entitled "How much do you spend on auto repairs?"

a. What is the expected number of cars per randomly selected household?

b. What is the standard deviation of the probability distribution for the number of cars per randomly sampled household?

c. If the poll selects 1000 households, how many cars are expected to be represented in the poll?

d. If the polling organization expects 1000 cars to be represented, how many households should be sampled?

e. Car maintenance costs average $250 per year, not counting gas and oil. How much would a randomly selected household expect to pay annually for car maintenance?

f. What is the standard deviation of the amount a randomly sampled household is to pay for car maintenance?

g. Use the information in Problem 5e to determine the total amount a random sample of 1000 families would expect to pay for car maintenance in a year. Do you think this expected value will be a good approximation to the real total for the 1000 families? (HINT: Find the standard deviation of the total amount 1000 families might have to pay for car maintenance and use that value in your answer.)

6. The following table shows the age distribution of residents of the United States for the years 1990 and 1996, according to the U.S. Bureau of the Census. The population figures are given in thousands. The columns labeled "proportion" show the proportions of the residents in each of the age categories.

	Age	1990	1990 Proportion	1996	1996 Proportion
1	Under 5	18,849	0.076	19,354	0.073
2	5 to 9	18,062	0.072	19,640	0.074
3	10 to 14	17,189	0.069	19,131	0.072
4	15 to 19	17,750	0.071	18,699	0.070
5	20 to 24	19,135	0.077	17,307	0.065
6	25 to 29	21,233	0.085	19,004	0.071
7	30 to 34	21,906	0.088	21,217	0.080
8	35 to 39	19,975	0.080	22,508	0.085
9	40 to 44	17,790	0.071	20,940	0.079
10	45 to 49	13,820	0.055	18,474	0.069
11	50 to 54	11,368	0.046	14,216	0.053
12	55 to 59	10,473	0.042	11,429	0.043
13	60 to 64	10,619	0.042	9,997	0.038
14	65 to 69	10,077	0.040	9,873	0.037
15	70 to 74	8,022	0.032	8,773	0.033
16	75 to 79	6,145	0.025	6,928	0.026
17	80 to 84	3,934	0.016	4,587	0.017
18	85 to 89	2,049	0.008	2,399	0.009
19	90 to 94	764	0.003	1,020	0.004
20	95 to 99	207	0.001	288	0.001
21	100 or more	37	0.000	58	0.000

a. Show appropriate plots of the age distribution for 1990 and the age distribution for 1996.

b. Discuss the key differences between the shapes of the two age distributions. What is the major change in the age distribution between 1990 and 1996?

c. Approximate the median age for the 1990 population. Do the same for the 1996 population. Compare the median ages.

d. Suppose the Gallup organization is to take a random sample of a large number of residents of the U.S. What can they expect as the mean age of those in their sample? How does this expected value compare to the median found above? How does this expected value compare to a similar expectation found for a sample taken in 1990?

e. Under the conditions described in Problem 6d, what can the Gallup organization expect as the standard deviation of the ages of the people who end up in their random sample? In this case, is the standard deviation a good description of a "typical" deviation from the mean age?

7. According to Census data, about 90% of the U.S. work force have at least a high-school education, about 57% have at least some college education, and about 29% have at least a bachelor's degree from a college or university. Suppose a typical worker without a high-school education earns about $14,000 per year, a typical high-school graduate makes about $20,000 a year, a typical worker with some college experience but not a bachelor's degree makes about $23,000 a year, and a typical worker with at least a bachelor's degree makes about $43,000 per year. Find the expected yearly income for a person randomly selected from the U.S. workforce. Explain why this expected value may be slightly different from the true mean income of the U.S. workforce.

Unit II

Sampling Distributions of Means and Proportions

LESSON 5

The Distribution of a Sample Mean

If you were to sample 100 families, what is the total number of children you would expect to see?

What is the distribution of potential values of the mean number of children per family in the sample of 100 families?

How will the distribution of potential values of the mean change with the sample size?

OBJECTIVE

Understand the behavior of the distribution of means from random samples.

What is the average summer daytime temperature for your town? What is the average age of a student in your class? What is the average number of points scored by your basketball team during the season? What is the average time it takes you to get to school in the morning? Averages used as summary or typical numerical values are all around us. When working with data, we have seen that the arithmetic average is called the mean. When working with a probability distribution, a possible model for data yet to come, the average is called the expected value.

Much of the data we see in designed studies, such as sample surveys, comes about through random samples from specific populations. Since these data are typically reported in summary form as means, it is important that we understand the behavior of sample means that arise from random sampling.

INVESTIGATE

The A. C. Nielsen Company samples households to collect data on TV-viewing habits. For some shows, the company is particularly interested in the number of children under the age of 18

who might be watching. Thus, it wants to make sure that there will be a reasonable number of children in its sample of households. One way to predict where this number might lie is to study the possible values of the mean number of children per sampled family in a typical random sample of families.

Something is known about the number of children in U.S. families, and that is the place to begin. The available population information comes in the form of the Census Bureau's distribution of children per family, as used in earlier lessons. The data are in the table below.

Number of Children	Percent of Families
0	51
1	20
2	19
3	7
4	3

An approximation to the expected number of children per family, as calculated in earlier lessons, is approximately 0.91.

From information gathered so far, Nielsen can tell something about the expected number of children per family in a random sample of n families. What other information does the company need?

Discussion and Practice

Nielsen wants to choose a sample size that produces a reasonable number of children with high probability.

1. If $n = 1000$, how many children would Nielsen expect to see in the sample of families? Suppose the company has a goal of seeing at least 1000 children in its survey. Do you think the probability of seeing more than 1000 children is high for a sample of 1000 families?
2. Suppose the sample size is increased to $n = 1200$. Will that increase the probability that Nielsen will achieve the goal of at least 1000 children in the sample? Will that probability change a great deal or very little?

In order to get specific answers to questions like those just posed, more information on the probability distribution of sample means from random samples must be developed; that is

the goal of this lesson. This new information will be discovered through simulation.

The Sampling Distribution of a Mean

If one family is randomly selected from the U.S. population, what is the probability that the family will contain no children? One child? Four children? Our first job is to design a simulation for random sampling that will preserve these probabilities, yet allow us to look at typical sample data on children per family. This investigation can be completed most efficiently if you work in small groups and then combine data for the class.

3. Designing and conducting the simulation

- **a.** Each group must have 100 plastic chips or small pieces of paper of equal size. Number the chips with integers 0 through 4, in the same percents given in the table for number of children per family, and place the chips into a box. You have now constructed a physical model of the probability distribution of the number of children per family. If you reach into the box of chips and randomly select one chip, what is the probability that it will show a 2? Is this the probability you want for modeling Nielsen's sampling process?
- **b.** Select a simulated sample of 10 households. Randomly draw a series of 10 chips from the box, but be careful to *replace* each sampled chip before the next one is drawn. Why is it important to do the sampling with replacement?
- **c.** What percent of the sample outcomes were *zeros*? What percent were *ones*? Is this approximately what you expected?
- **d.** Calculate the mean number of children per household found in your random sample of 10 households.
- **e.** Repeat the process for three more samples of 10 households each. Record and save the sample data and the values of the means. Your group should now have four samples.
- **f.** Collect the sample means produced by all the groups in the class. Plot the collection of sample means on a dotplot or stemplot so that the shape of the distribution can be seen. Comment on this shape. Does it differ from the

shape of the population distribution of children per family? How does it differ?

4. Changing the sample size

   a. Repeat the simulation of Problem 3 for samples of 20 households each. This can be accomplished by making two pairs of the size-10 samples already selected, yielding two size-20 samples. Collect the sample means produced by all groups in the class, plot the means on a dotplot or stem-and-leaf plot, and comment on the shape.

   b. Repeat the simulation of Problem 3 for samples of 40 households each. This can be accomplished by combining the two size-20 samples already available in each group. Collect the sample means produced by all groups in the class, plot the means on a dotplot or stem-and-leaf plot, and comment on the shape.

   c. Repeat the simulation of Problem 3 for samples of 80 households each. This can be accomplished by combining the 40 data points from your group with 40 from another group. Make sure each group's data is used only once. Collect the sample means produced by the class, plot the means on a dotplot or stem-and-leaf plot, and comment on the shape.

   d. The distribution of possible values of the sample mean is called a ***sampling distribution***. Study the plots of sampling distributions for size-10, size-20, size-40, and size-80 samples, and comment on

      i. the shapes of the sampling distributions of sample means.

      ii. the centers of the sampling distributions of sample means.

      iii. the variation in the sampling distributions of sample means.

5. Computing summary statistics

   a. Calculate the means and the standard deviations for each of the simulated sampling distributions produced above. That is, use the original sets of sample means generated by the class in the simulations to calculate the mean and the standard deviation of the sample means within each sample size.

b. How do the calculated means of the simulated sampling distributions compare to the expected number of children per family (the mean of the population) calculated from the population distribution to be $\mu = 0.91$? (The symbol μ used for the population mean is the Greek mu, or m.) Make a general statement about how the means of the sampling distributions relate to the mean of the population from which the samples were selected.

c. How do the calculated standard deviations of the simulated sampling distributions compare to the standard deviation of the population, calculated to be $\sigma = 1.114$? (The symbol σ used for the population standard deviation is the Greek sigma, or s.) Do you see a pattern developing in how the standard deviations of the sampling distributions relate to the sample sizes?

d. The precise relationship between the standard deviation of a sampling distribution for means and the sample size is difficult to see intuitively, so we'll provide some help. We denote the population standard deviation by σ and label the standard deviation of a sampling distribution of sample means by *SD*(mean). You have noticed that *SD*(mean) decreases as the sample size n increases. Mathematical theory of statistics says that the precise relationship among these quantities is given by

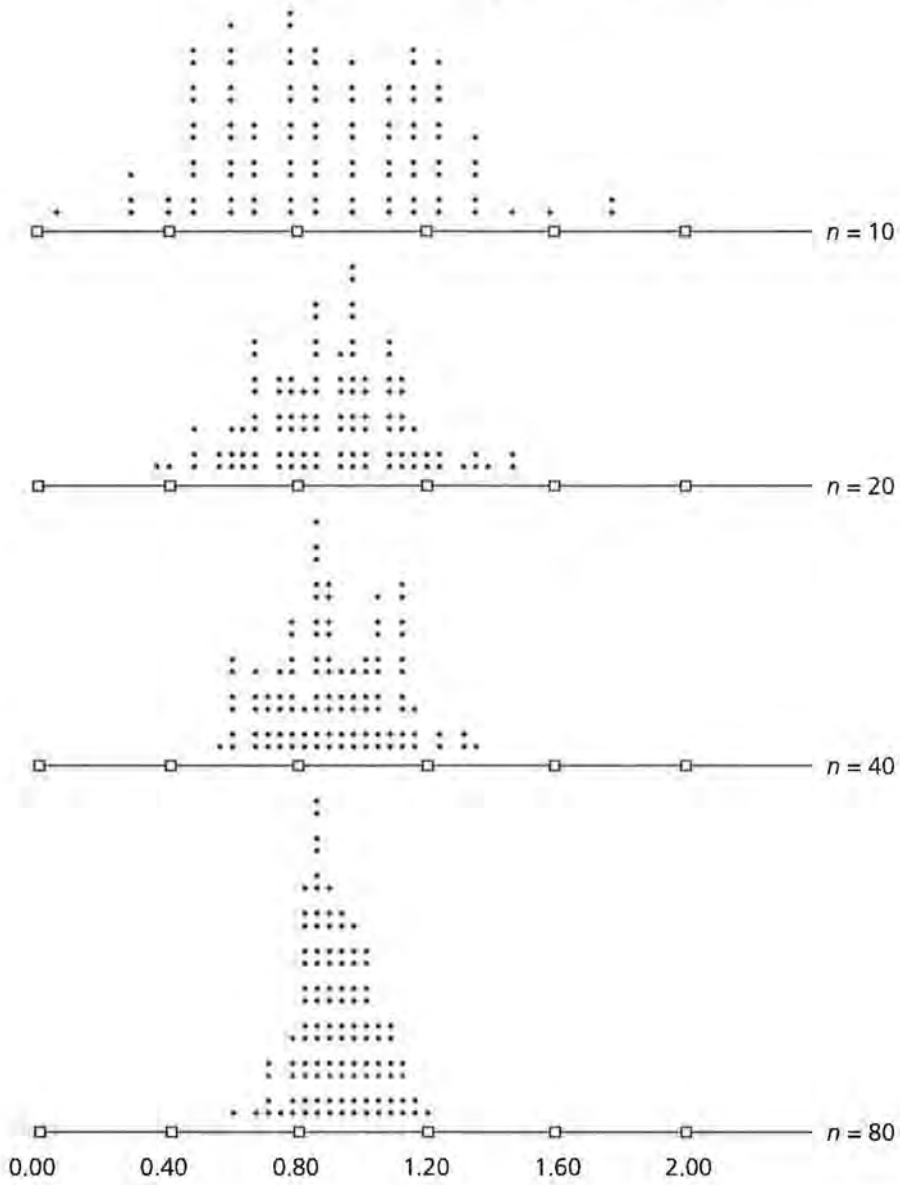
$$SD(\text{mean}) = \frac{\sigma}{\sqrt{n}}$$

For $\sigma = 1.114$, as it is for the population of number of children per household, calculate the theoretical *SD*(mean) for sample sizes of 10, 20, 40, and 80. Compare these theoretical values to the observed standard deviations of the sampling distributions calculated in Problem 5a.

PATTERNS

To show that the patterns generated in sampling distributions simulated by your class are not accidental, and to show a slightly larger simulation, we have produced a computerized version of the simulation outlined above. In this case, we have calculated the means from 100 random samples in each simulated sampling distribution. The results are shown in the dotplots that follow.

Sampling Distributions of Sample Means



	No. Samples	Mean	Standard Deviation
$n = 10$	100	0.9140	0.3426
$n = 20$	100	0.9320	0.2397
$n = 40$	100	0.8970	0.1891
$n = 80$	100	0.9165	0.1188

Notice three things:

- The sampling distributions all center around the population mean of 0.91.
- The standard deviations of the sampling distributions get smaller as the sample size increases.

- The sampling distributions tend to have somewhat symmetric mound shapes.

To be more specific about the second point, let's compare the observed standard deviations of the sampling distributions to the theoretical values that should be generated:

n	*SD* observed	$SD(\text{mean}) = \frac{\sigma}{\sqrt{n}}$
10	0.3426	0.3523
20	0.2397	0.2491
40	0.1891	0.1761
80	0.1188	0.12457

It appears that the theoretical rule for relating the standard deviation of the sampling distribution to the population standard deviation and the sample size works well. We will return to the point about the symmetric, mound-shaped distribution in the next lesson.

Practice and Applications

6. Suppose the Nielsen Company wants a sample of families containing at least 30 children in all. Is this highly likely with a random sample of the size given?

a. $n = 10$

b. $n = 20$

c. $n = 40$

d. $n = 80$

Explain how to use the simulated sampling distributions on page 40 to answer this question.

7. Suppose the Nielsen company is to select a random sample of 1000 families.

a. Describe the distribution of potential values of the sample mean number of children per family. The description should include a statement about the center and spread of the distribution of potential values.

b. If Nielsen wants to see at least 1000 children in the sample, what would the mean number of children per family have to equal or exceed? Do you think it is likely that the sample of 1000 households will produce at least 1000 children? (HINT: It is unusual for a data value to

lie more than two standard deviations from the mean of the distribution from which it was selected.)

c. Suppose Nielsen changes to a random sample of 1200 households. Does this dramatically improve the chance of seeing at least 1000 children in the sampled households? Explain.

8. The people using an elevator in an office building have an average weight of approximately 150 pounds and a standard deviation of weights of approximately 10 pounds. The elevator is designed for a 2000-pound weight maximum. This maximum can be exceeded on occasion, but should not be exceeded on a regular basis. Your job is to post a sign in the elevator stating the maximum number of people for safe use. Keep in mind that it is inefficient to make this number too small, but dangerous to make it too large. What number would you use for maximum occupancy? Explain your reasoning.

9. A call-in radio show collects callers' opinions on the number of days students should be in school during a year. The mean number for 500 callers was 195 days. The radio show then announces that this mean should be close to the mean one would obtain if all residents of the community were asked this question. What is wrong with this reasoning?

10. What happens to the mean of the sampling distributions as the sample size increases, everything else remaining fixed? How does the mean of the sampling distributions compare to the mean of the population from which the samples were selected?

11. What happens to the standard deviation of the sampling distributions as the sample size increases, everything else remaining fixed? How does the standard deviation of each of the sampling distributions compare with the standard deviation of the population from which the samples were selected?

SUMMARY

Means, or averages, are one of the most common summary statistics used to describe data. Thus, to make inferences from data we must understand how means of random samples behave. The distribution of potential values of the sample mean to be produced by a random sample from a fixed population is called a ***sampling distribution***. Sampling distributions for

means have three important properties:

- The sampling distributions all center around the population mean.
- The standard deviations of the sampling distributions get smaller as the sample size increases, and this can be predicted by the rule

$$SD(\text{mean}) = \frac{\sigma}{\sqrt{n}}$$

- The sampling distributions tend to be symmetric and mound-shaped.

The first two properties were investigated in this lesson; the third is the subject of the next lesson. This result is commonly known as the "Central Limit Theorem."

The Normal Distribution

What is the chance that your school's mean weekly earnings from recycling aluminum cans during the fall will exceed $130?

Do the probability distributions of potential values of a sample mean always have nearly the same shape?

How can you make use of a common model for distributions of means?

OBJECTIVES

Understand the basic properties of the normal distribution.

See the usefulness of the normal distribution as a model for sampling distributions.

Means are widely used statistical summaries, and decisions based on means can be more enlightened if decision-makers understand the behavior of sample means from random samples. Suppose records show that the weekly amounts your school earns on recycling aluminum cans has a mean of $120 and a standard deviation of $8. During a sixteen-week period in the fall, what is the chance that the mean weekly earnings will exceed $130? $124? $100? In Lesson 5, you used simulation to answer questions similar to these. But it is cumbersome and time-consuming to conduct a simulation every time you want to answer a question about a potential value of a mean. It would be helpful to have a ***model*** for the behavior of sample means which would give quick approximate answers to the many questions that arise about sample means. Such a model is the ***normal distribution.***

For the A. C. Nielsen Company, it may be important to know even more details than provided in Lesson 5 about the possible values of mean number of children per sampled family. One such question might be, "What is the chance of having fewer than 1000 children in a sample of 1200 families?" Relating the

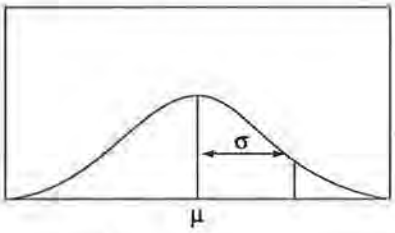
sampling distributions discovered in Lesson 5 to a theoretical model for such distributions—the normal distribution—can help provide answers to such questions.

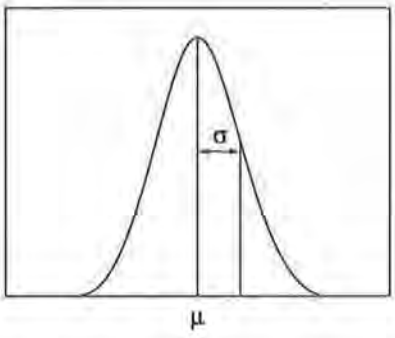
INVESTIGATE

Consider, once again, the plots constructed in Lesson 5 (on page 40) that show the probabilistic behavior of sample means for various sample sizes. These curves look fairly symmetric and mound-shaped. Such mound-shaped, symmetric distributions are seen very often in the practice of plotting data. In fact, they are seen so often that such a curve is called "normal," and a theoretical model for this curve has been studied extensively.

Normal curves have two key measurements that determine their location and shape. One, the location of the center of the curve on the real-number line, is the mean, usually denoted by m. The other, the measure of spread, or width, of the curve, is the standard deviation, usually denoted by s. Pictures of two normal curves with different standard deviations are shown below. Note that standard deviation measures variability. A careful look at these pictures shows the standard deviation to be half of the curve about $\frac{2}{3}$ of the way up the line of symmetry, the vertical line through the mean.

Normal Distributions





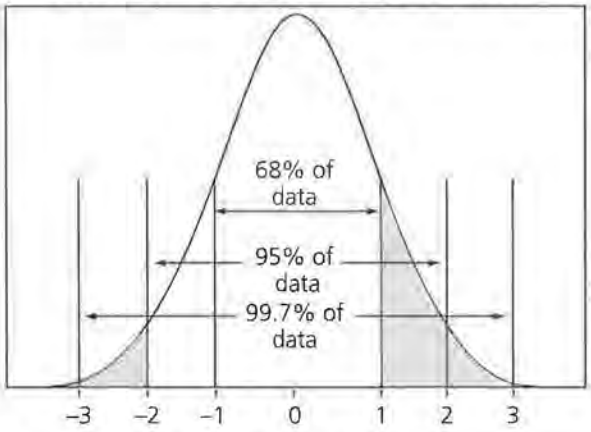
The area under the curve over an interval on the horizontal axis represents the percent of the data that fall into that interval. These intervals can be located in terms of the mean and

standard deviation of the distribution of data. For the normal curve, it is common practice to describe the distribution in terms of intervals that are symmetric about the mean. For the normal distribution,

> the interval $\mu \pm \sigma$ contains about 68% of the data,
> the interval $\mu \pm 2\sigma$ contains about 95% of the data,
> and the interval $\mu \pm 3\sigma$ contains about 99.7% of the data.

These intervals and their respective areas are shown in the graph that follows. The scale on the horizontal axis is in terms of standard-deviation units. A point at 1 is one standard deviation above the mean. A point at –2 is two standard deviations below the mean.

Areas Under a Normal Distribution



Discussion and Practice

1. For a distribution of data that can be represented by a normal curve,
   a. what percent of the data is below the mean?
   b. what percent of the data is more than one standard deviation above the mean?
   c. what percent of the data is between one standard deviation below the mean and two standard deviations above the mean?
   d. what percent of the data is more than two standard deviations away from the mean?
2. Does the population distribution of number of children per family (See Lesson 1 or the Assessment for Lessons 1–4.) look "normal"? Explain your reasoning.

3. Recall other distributions of data you have seen in these lessons or other places. Describe at least two other data sets that look normal to you. Explain your reasoning.

The Normal Distribution as a Model

The task at hand is to relate the normal-looking sampling distributions for the sample mean, found in Lesson 5, to the normal distribution. The normal distribution, we discovered above, depends upon two constants, a mean and a standard deviation. Each of the sampling distributions seems to have approximately the same mean, and it is close to the population mean, or expected value, of 0.91 in the case of number of children per family. Therefore, we can take $\mu = 0.91$ in our theoretical normal model.

What about the standard deviation for the normal model? The standard deviations within the sampling distributions decrease as n increases, so each sample size generates a slightly different sampling distribution. These standard deviations are related to the population standard deviation by the equation

$$SD(\text{mean}) = \frac{\sigma}{\sqrt{n}}$$

where σ denotes the theoretical standard deviation of the underlying population. This relationship was explored in Lesson 5, with $\sigma = 1.114$ for the distribution of number of children per family.

4. Summarize the above information in a concise statement about the approximate sampling distribution of sample means for random samples of size n from a fixed population, by answering the questions below.
   a. What is the shape of the sampling distribution?
   b. Where is the center of the sampling distribution?
   c. What is the standard deviation of the sampling distribution?
5. Rewrite the basic relative-frequency rules for the normal distribution, preceding Problem 1 above, in terms of sampling distributions for sample means.
6. Look carefully at the distributions on page 40 of Lesson 5. What fraction of the observed sample means lie within two standard deviations of the mean of their distribution in each of the four cases? Do these fractions agree with what the normal distribution would predict?

Another Simulated Sampling Distribution

Do you think the normal distribution would apply to sampling distributions of means for other population distributions? After all, we have based most of our discussion on a single underlying distribution–that of the number of children in American families.

7. This simulation will involve the use of random digits. You will need a random-digit table or a calculator that generates random digits. The simulation is completed most efficiently by working in groups.

   a. Work with your group to select 50 sets of ten random digits each from a random-number table, or random-number generator in your calculator. Compute the mean of each set of ten digits.

   b. Plot the 50 means on a dotplot. Describe the shape. Does the plot look normal?

   c. Compute the mean and the standard deviation of the set of 50 sample means.

   d. Random digits take on integer values from 0 to 9 with equal probability. Using this fact, compute the expected value and the standard deviation as an expected value for the theoretical distribution of random numbers.

   e. How does the observed mean of the simulated sampling distribution compare with the theoretical expected value? How does the observed standard deviation of the simulated sampling distribution compare with the theoretical standard deviation? Does it look as if the rules developed above for the sampling distribution for means apply in this case?

8. Consider the seeming generality of the normal model for sampling distributions of means, for a random sample of size n from a population with mean μ and standard deviation σ. Give the probability that the following inequality holds:

$$|\bar{x} - \mu| \leq \frac{2\sigma}{\sqrt{n}}$$

where $\bar{x}$ denotes the sample mean. (NOTE: This statement is the same as, "What is the probability that the sample mean will be within two standard deviations of its population mean?")

Practice and Applications

9. For the random samples of families that the Nielsen Company could select, what interval should contain the middle 95% of the sample means on number of children per family for samples of each size?

a. 25

b. 100

c. 1000

d. 4000, which is the approximate real sample size for Nielsen ratings

10. What interval should contain the middle 95% of possible values for the total number of children in random samples of each size?

a. 25

b. 100

c. 1000

d. 4000

11. How large should the sample size be to insure that the total number of children in the sample is below 1000 with probability of only about 0.025?

12. Below is a restatement of the question at the beginning of this lesson. Read the statement and approximate the chances asked about. Do you need to make any assumptions for your answers to be valid?

Suppose records show that the weekly amounts your school earns on recycling aluminum cans has a mean of $120 and a standard deviation of $8. In the sixteen-week period of the fall term, what is the chance that the mean weekly earnings will exceed $130? $124? $100?

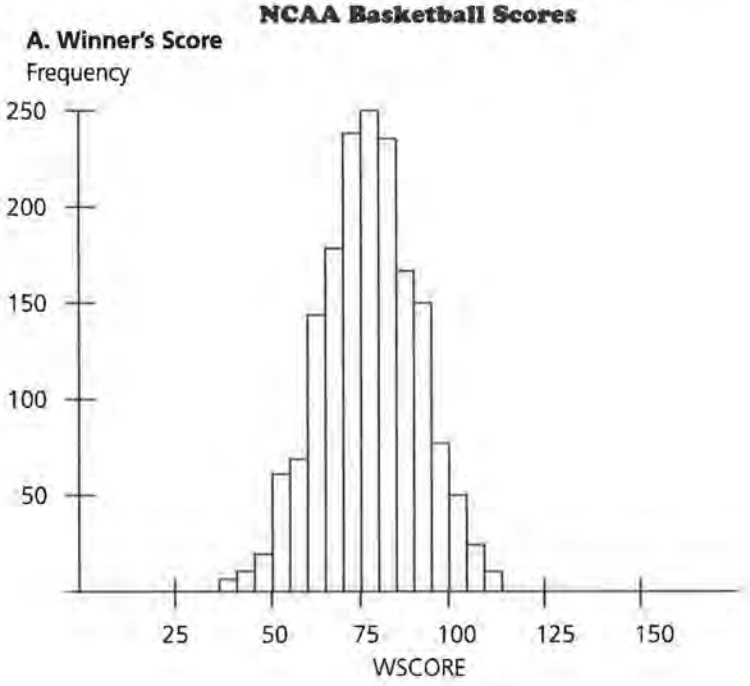
13. Here is a restatement of Problem 8 from Lesson 5:

The people using an elevator in an office building have an average weight of approximately 150 pounds, and a standard deviation of weights of approximately 10 pounds. The elevator is designed for a 2000-pound weight maximum. This maximum can be exceeded on occasion, but should not be exceeded on a regular basis. Your job is to post a sign in the elevator stating the maximum number of people for safe use. Keep in mind that it is inefficient to make this

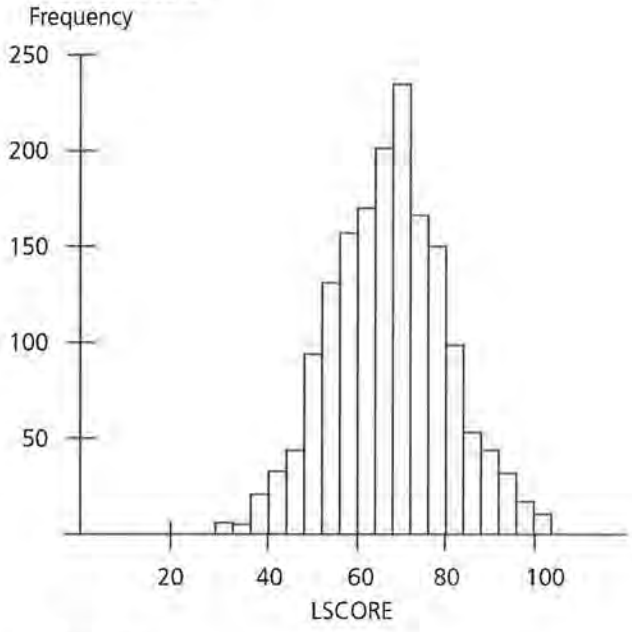
number too small, but dangerous to make it too large. What number would you use for maximum occupancy? Explain your reasoning.

What is the approximate probability that the weight limit will be exceeded if the number of people on the elevator is 14? 13? 12?

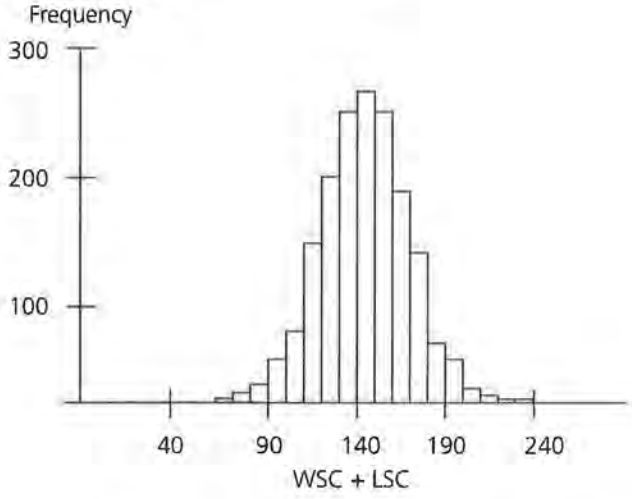
14. The normal distribution sometimes works well as a model of the distribution of the population itself. If the population has a normal distribution, then the sampling distribution for a sample mean will be normally distributed for any sample size, even a sample size of 1. The three histograms that follow show basketball scores from all NCAA college playoff games between 1939 and 1995. Plot A is based on the score of the winner, plot B on the score of the loser, and plot C on the total points scored in the game.



B. Loser's Score



C. Total Score



a. Does the normal distribution seem like a good model for these distributions? Which of the three do you think is least normal-looking? Explain.

b. Approximate the mean and standard deviation for each of the three distributions.

c. Find an interval which includes approximately 95% of the scores of the winning team.

d. Would you expect to see the total score in an NCAA playoff game go above 180 very often? Explain.

SUMMARY

Sampling distributions for sample means, the distributions of potential values of a sample mean, are used widely in making decisions with data. The ***normal distribution*** provides a good ***model*** for these sampling distributions as long as the samples are randomly selected from a fixed population. The normal distribution is characterized by its symmetric mound shape, mean, and standard deviation. From knowledge of the mean and standard deviation, relative frequencies within intervals to either side of the mean can be found.

LESSON 7

The Distribution of a Sample Proportion

How many high-school graduates would you expect to see in a random sample of 1000 adults?

Is there a good chance that a sample of 1000 adults could produce over 900 high-school graduates?

Is the distribution of potential values of a sample proportion similar to the distribution of potential values for a sample mean?

What percent of the students in your school like the food in the cafeteria? What proportion of your income do you spend on food and entertainment? What fraction of the residents of your town own their homes? Data are often summarized by reporting a percent or a proportion for one or more categories involved. In order to make intelligent decisions based on data reported this way, we must understand the behavior of proportions that arise from random samples.

OBJECTIVES

Gain experience working with proportions as summaries of data.

Develop sampling distributions for sample proportions.

Discover the meaning of margin of error in surveys.

INVESTIGATE

The Sampling Distribution of a Proportion

First, let's be clear about the basic terminology. Suppose 20 students are asked to report their favorite food and 12 say "pizza." The ***proportion*** reporting pizza is $\frac{12}{20} = 0.60$ and the ***percent*** reporting pizza is 60%. Most of the problems in this lesson will deal with proportions, which will be numbers between 0 and 1.

Discussion and Practice

Thinking About Survey Results

1. It is estimated that about 60% of automobile drivers use seat belts. Suppose your class is to conduct a survey of 40 randomly selected drivers. Think about the following questions. Try to arrive at reasonable answers based upon your current knowledge and your intuition. Do not spend a lot of time with calculations.

   a. How many drivers would you expect to be using seat belts?

   b. What is the chance that more than 30 drivers will be using seat belts?

   c. Would it be quite unusual to find fewer than 10 of the drivers wearing seat belts?

Now, we will develop the tools we need to answer these questions more precisely.

An intuitive "feel" for how sample percents behave in random sampling can be developed through simulation. Let Y denote the number of successes in a random sample of size n from a population in which the probability of success on any one sample selection is given by p. This section will concentrate on properties of the sample proportion Y/n. Y/n can be changed into a percent by multiplying by 100, but we will work most directly with the decimal form. This investigation can be completed most efficiently if you work in pairs.

Designing and Conducting the Simulation

2. Assume that $p = 0.6$ for this study.

   a. Find a device that will generate an outcome that has probability equal to 0.6 of occurring. A random-number table, a calculator that generates random numbers, or slips of paper may be used.

   b. Let the sample size be $n = 10$. Generate ten outcomes by the device agreed upon in Problem 2a and count the successes among the ten outcomes. Divide the number of successes by 10 to obtain the proportion of successes. That is the sample proportion that will be recorded.

   c. Repeat Problem 2b three more times so that your group has four different samples of size 10 each and four observed values of the sample proportion.

d. Combine your values of the sample proportion with those from the rest of the class. Plot the sample proportions on a dotplot or a stem-and-leaf plot.

e. Based on the results of the simulation, find approximate values for each probability.

i. $P(0.5 < \frac{Y}{n} < 0.7)$

ii. $P(0.4 < \frac{Y}{n} < 0.8)$

iii. $P(\frac{Y}{n} > 0.8)$

f. Calculate the mean and the standard deviation of the simulated distribution of sample proportions. Keep these for future reference.

g. From your dotplot or stem-and-leaf plot, describe the distribution of the possible values of $\frac{Y}{n}$ for samples of size 10, when the true value of p is 0.6. NOTE: This distribution is called a simulated sampling distribution of the proportion $\frac{Y}{n}$.

Changing the Sample Size

3. a. Repeat the simulation of Problem 2 for samples of size $n = 20$. Keep p fixed at 0.6. To save time, two of the size-10 samples may be combined to obtain a size-20 sample. Complete all parts of Problem 2.

b. Repeat the simulation of Problem 2 for samples of size $n = 40$. Keep p fixed at 0.6. To save time, two of the size-20 samples may be combined to obtain a size-40 sample. Complete all parts of Problem 2.

c. Compare the shapes of the simulated sampling distributions for samples of size 10, 20, and 40.

d. Compare the means of the simulated sampling distributions for samples of size 10, 20, and 40. How close are these means to the true value of p?

e. Compare the standard deviations of the simulated sampling distributions for samples of size 10, 20, and 40. A theoretical result in probability states that the standard deviation of sample proportions should be related to the true p and the sample size by the following rule:

$$SD(\text{proportion}) = \sqrt{\frac{p(1-p)}{n}}$$

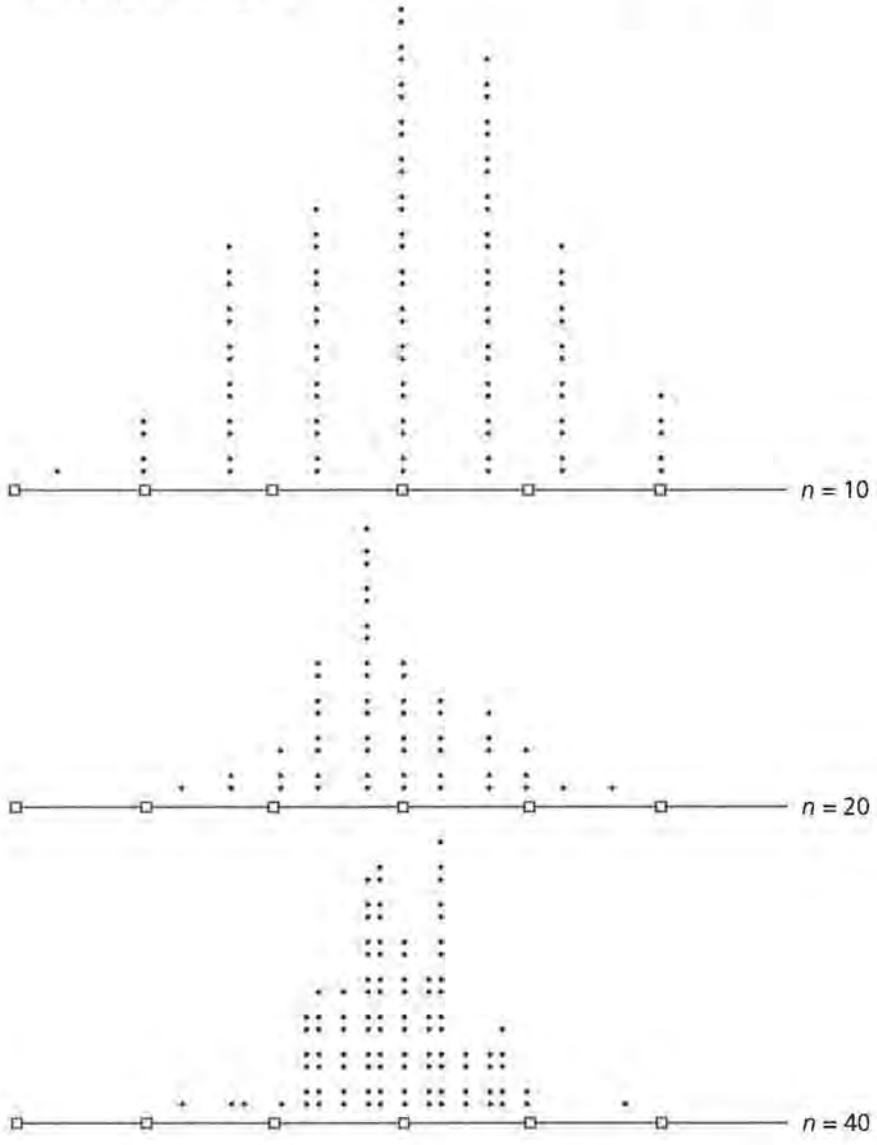
Does this rule appear to hold for the sampling distributions generated in this investigation?

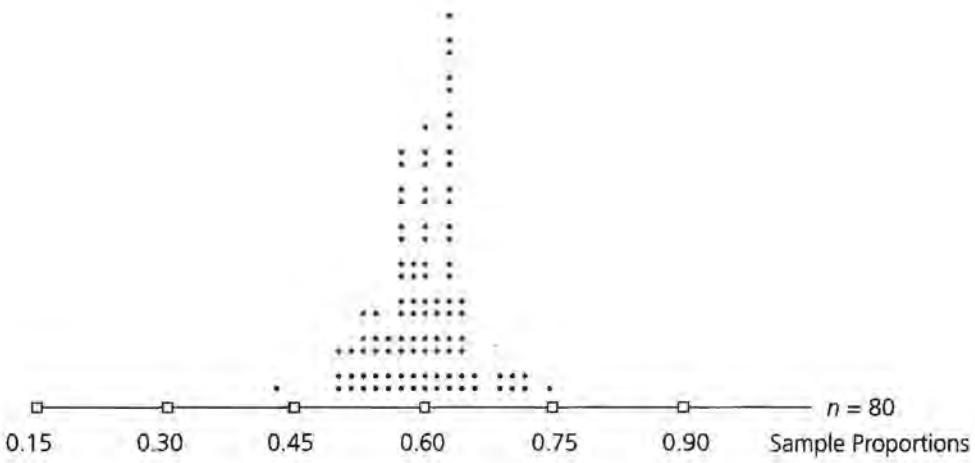
Summary of Sampling Distributions for Sample Proportions

4. Your plots may look something like those shown below. These were generated by computer for the same cases described above, with the addition of a sampling distribution for samples of size 80.

Sampling Distributions of Sample Proportions

Each dot represents 2 points.





	Mean	Median	Standard Deviation
$n = 10$	0.6070	0.6000	0.1552
$n = 20$	0.5790	0.5500	0.0962
$n = 40$	0.5910	0.5750	0.0838
$n = 80$	0.5976	0.6000	0.0540

a. What are the main similarities and differences among these plots?

b. Where are the centers of these distributions? Are they all close to 0.6?

c. The observed standard deviations for the four sampling distributions are given following the four plots. How do these compare with the theoretical standard deviations calculated by the following rule?

$$SD(\text{proportion}) = \sqrt{\frac{p(1-p)}{n}}$$

Does the rule appear to work reasonably well?

The Normal Model for Proportions

From your simulation and the one seen in Problem 4, it appears that the sampling distributions for proportions are somewhat mound-shaped and symmetric. Thus, they could be modeled by the normal distribution with mean p and standard deviation

$$\sqrt{\frac{p(1-p)}{n}}$$

The following questions are intended to amplify this normality of the sampling distributions for proportions.

5. Refer to the four sampling distributions in Problem 4. Remember, in each case $p = 0.6$.

   a. Find the fraction of the simulated values of $\frac{Y}{n}$ that were within two standard deviations of p. Do this separately for each of the distributions in the figure.

   b. Symbolically, the fractions found in Problem 5a are estimates of

$$P\left[\left|\frac{Y}{n} - p\right| \leq 2\sqrt{\frac{p(1-p)}{n}}\right]$$

   For each of the cases $n = 10$, $n = 20$, $n = 40$, and $n = 80$, are these estimated probabilities close to each other?

   c. Theoretically, about 95% of the observed values of a sample proportion will be within two standard deviations of the expected value of that proportion. This implies that it is quite likely, in any one sample, to obtain a sample proportion of successes that is less than two standard deviations from the "true" proportion of successes in the population. Do the results of Problem 5b appear to bear this out?

Practice and Applications

6. Look at the results of some recent polls as reported in the media—newspapers, news magazines, and so on.

   a. Do the results of these polls tend to be reported in terms of counts, that is, number of "successes," or in terms of percents?

   b. Discuss reasons why percents might be better than counts as a way to summarize data from polls.

   c. Discuss aspects of the reporting on polls that you would like to see improved. Base the discussion on printed articles you have read.

7. If you look carefully at articles from the media on opinion polls, which are sample surveys, you may observe that many of them contain a statement similar to one of the following:

   "The margin of error in these percents is 3%."

   "The sampling error in any one reported percent is 2.5%."

   What do you think they mean by these phrases? (HINT: Look again at Problem 5.)

8. Speculate as to what would happen to the shapes, means, and standard deviations of the simulated sampling distributions of in $\frac{Y}{n}$ Problem 4 if p were changed to 0.5. What if p were changed to 0.8?

9. According to the U.S. Bureau of the Census, about 80% of U.S. residents over the age of 25 are high-school graduates. A Gallup poll is to be conducted among those over the age of 25—we'll call these "adults" for now—on issues surrounding education. So the Gallup organization wants to be sure there will be adequate numbers of both high- school graduates and non-graduates in the resulting random sample of adults.

   a. What is the probability that a randomly selected adult is a high-school graduate?

   b. If 1000 adults are randomly sampled, how many high-school graduates would you expect to see?

   c. If Gallup expects to have 400 non-graduates in the sample, how many adults should be randomly selected?

   d. Can Gallup be assured of at least 500 non-graduates in a poll of any fixed size?

   e. If Gallup randomly samples 400 adults, is it likely that the poll will result in more than 335 high-school graduates? Explain.

10. Here is a restatement of the first set of questions posed in this lesson. Answer them as specifically as you can with the information gained in this lesson.

    It is estimated that about 60% of automobile drivers use seat belts. Suppose your class is to conduct a survey of 40 randomly selected drivers. Think about the following questions. Try to arrive at reasonable answers based upon your current knowledge and your intuition. Do not spend a lot of time with calculations.

    a. *How many drivers would you expect to be using seat belts?*

    b. *What is the chance that more than 30 drivers will be using seat belts?*

    c. *Would it be quite unusual to find fewer than 10 of the drivers wearing seat belts?*

SUMMARY

Sample ***proportions*** or percents are used often to summarize data on categorical variables, like gender, ethnic group, attitude on public issues, and health condition. In order to make decisions based on this kind of data, we must know something about the anticipated behavior of sample proportions from random samples. What we have discovered is that sample proportions have approximately normal sampling distributions, with mean equal to the true population proportion p for the characteristic being studied and standard deviation given by the following rule:

$$\text{SD(proportion)} = \sqrt{\frac{p(1-p)}{n}}$$

Lessons 5–7

1. Explain what a sampling distribution for a sample mean is and how it differs from the underlying population distribution.

2. According to the U.S. Bureau of the Census, the number of people per household averages 2.58 and the standard deviation of the number of people per household is 1.39. (NOTE: See Lesson 2 for the actual population distribution for this variable.) Describe the sampling distribution of the sample mean for random samples from this distribution for each sample size.

   a. $n = 5$

   b. $n = 100$

   c. $n = 1000$

   Your description should include sketches of the sampling distributions.

3. Based on the information in Problem 2 on the distribution of household sizes, how large a sample size is needed to guarantee that the standard deviation of the distribution of possible sample means is less than 0.10?

4. Gallup is conducting a poll of American households. A typical sample size is 1200 households. For a sample of 1200 households, what interval should contain the middle 68% of the possible values for the sample mean?

5. What sample size should Gallup use if it is desired to have a total of at least 2000 people in the sample with probability about 0.975?

6. Suppose the distribution of household size in the U.S. has summary statistics as given in Problem 2. For a sample of 1200 households, the distance between the sample mean and the population mean should still be less than K with probability about 0.95. Find K. Did you make any assumptions in the process of finding K? What are they?

7. A random sample of n students in your school is to be selected to estimate the proportion p of students who will be requesting parking spaces for next year. Describe the shape, center and spread of the distribution of possible values of the sample proportion that could result from this survey.

8. About 60% of drivers wear seat belts. In a random sample of 100 drivers, what interval should contain the middle 95% of the possible sample proportions of seat belt users?

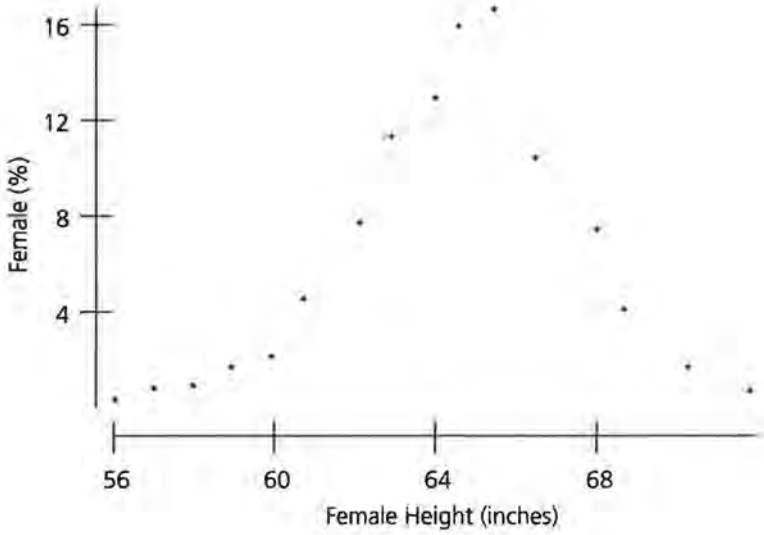
9. From prior studies we assume that the percent of drivers wearing seat belts is around 60%, but we do not know for sure. A new survey is commissioned to study this issue. What sample size will guarantee a margin of error no larger than 0.04?

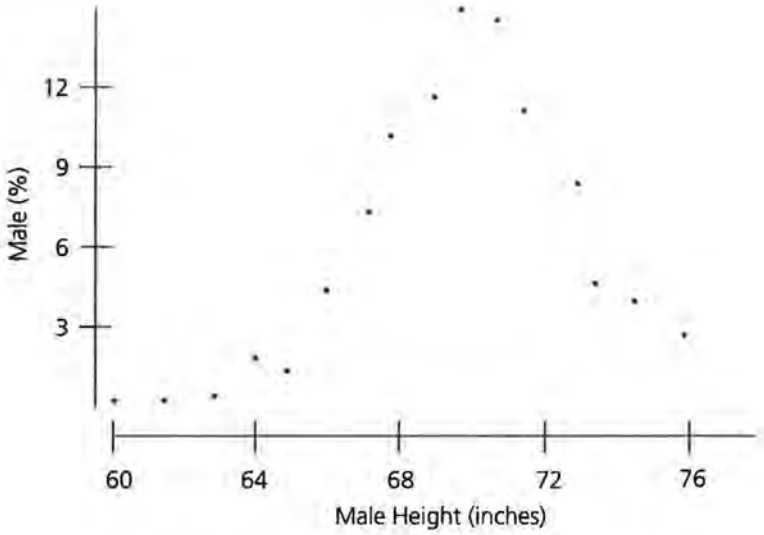
10. About 30% of the residents of the U.S. are without health insurance.

    a. If a poll of 500 residents is conducted, with random sampling, would it be unusual to find over 40% without health insurance? Explain.

    b. In a poll of 500 residents, suppose 42% were found to be without health insurance. Discuss various possible reasons for this unusually large sample percent.

11. The plots on the next page show the percent of U.S. adult females and adult males having heights that would round off to the integer values given on the horizontal axes. The female percents are given for heights from 56 through 71 inches and the male percents are given for heights from 61 through 76 inches.





a. What are the approximate mean and standard deviation of female heights?

b. What are the approximate mean and standard deviation of male heights?

c. Find a female height that would be exceeded by only about 16% of the adult females in the U.S.

d. Find a male height that has probability 0.84 of being exceeded by the height of a randomly select adult male from the U.S.

Two Useful Distributions

LESSON 8

The Binomial Distribution

In a random sample of four adults, what is the probability of sampling three high-school graduates?

How do small sample distributions differ from large sample distributions?

Of what use are mathematical models for calculating probabilities?

OBJECTIVES

Understand the basic properties of the binomial distribution.

Use the binomial distribution as a model for certain types of counts.

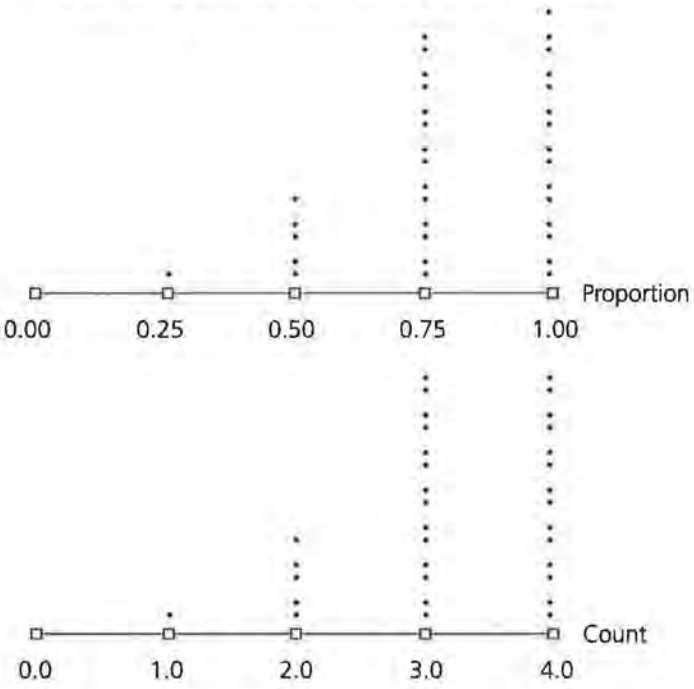
You learned in Lesson 7 that sample proportions tend to have normal sampling distributions; so the normal probability model works well for approximating probabilities associated with sample proportions. However, the normal approximations work well only in cases for which the sample size is large and the population proportion p is not close to zero or 1. For small samples or for cases in which p is far from 0.5, we need a more precise model.

INVESTIGATE

Sampling Distributions

It is reported that about 80% of adults over 25 in the U.S. are high-school graduates. Suppose a Gallup poll randomly selects only four adults from a small city for which this population percent is thought to hold. The first plot that follows shows a simulated sampling distribution for the sample proportion of high-school graduates in such a sample. The second plot shows the simulated distribution of the actual count of the number of high school graduates per sample of size 4.

Sampling Distributions for a Small Sample



Discussion and Practice

1. Describe the shapes of the two distributions plotted. Do they look as if they could be approximated well by normal distributions?

We will not attempt to give a general rule as to when the normal distribution is an appropriate approximation to the sampling distribution for a proportion; when in doubt, set up a small simulation to see if you think the normal model will work well.

A Random Variable for a Categorical Outcome

In the Gallup poll mentioned above, the outcome for any one sampled person is either "high-school graduate" or "not a high-school graduate." What can we do with these categorical outcomes to change them into numerical ones? One way to make the transition from categorical to numerical is to assign a 1 if we see the characteristic we are looking for and to assign a 0 if we do not see that characteristic. We might, for example, define "high-school graduate" to be of primary interest in the Gallup poll. Thus, for the first person sampled, we would record a 1 if that person is a high-school graduate and a 0 if not. Symbolically, we could write $X_1 = 1$ if sample person 1 is a

high-school graduate and $X_1 = 0$ if sample person 1 is not a high-school graduate.

The random variable X_1, then, completely describes the outcomes of interest for the first person sampled. The probability distribution for X_1 can be written as

X_1	$P(X_1)$
0	$1-p$
1	p

where p represents the probability of selecting an item with the characteristic of interest. In the high school graduate example, $p = 0.8$.

2. Use the random variable X_1 described above.

a. Find the expected value of X_1. How does it relate to the probability of sampling a high-school graduate?

b. Find the standard deviation of X_1. Write this as a function of p.

3. Suppose X_2 represents the outcome for a second adult sampled from the city, coded in the same way as X_1. Give the probability distribution for X_2. (HINT: The second adult is selected independently from the first.)

The Binomial Distribution

We are ready to develop a formula for the probability of obtaining a number Y of high-school graduates in a sample of n adults from a city in which p is the proportion of adults which are, in fact, high-school graduates. We begin by letting $n = 2$. In this case, Y can equal 0, 1, or 2; there are no other possibilities. Y is 0 if neither of the selected adults is a high-school graduate. Symbolically,

$$P(Y = 0) = P(X_1 = 0 \text{ and } X_2 = 0) = P(X_1 = 0) \cdot P(X_2 = 0) = (1-p)^2.$$

Multiplication is valid here since you are looking at the intersection of two independent events. For the sake of convenience in writing, let $(1 - p) = q$. Y is 1 if exactly one of the adults is a high-school graduate. Symbolically,

$$\begin{aligned} P(Y = 1) &= P((X_1 = 1 \text{ and } X_2 = 0) \text{ or } (X_1 = 0 \text{ and } X_2 = 1)) \\ &= P(X_1 = 1 \text{ and } X_2 = 0) + P(X_1 = 0 \text{ and } X_2 = 1) \\ &= P(X_1 = 1) \cdot P(X_2 = 0) + P(X_1 = 0) \cdot P(X_2 = 1) \\ &= pq + qp = 2pq. \end{aligned}$$

Why can we add probabilities on the second line?

Finally, Y is 2 if both adults are high-school graduates. Symbolically,

$$P(Y = 2) = P(X_1 = 1 \text{ and } X_2 = 1) = p^2.$$

By combining the above results, you can represent the probability distribution for Y in the case $n = 2$ with the following table.

Y	$P(Y)$
0	q^2
1	$2pq$
2	p^2

4. Show that the probability distribution of Y for the sample size $n = 3$ is given by the following table.

Y	$P(Y)$
0	q^3
1	$3pq^2$
2	$3p^2q$
3	p^3

From the probability distributions, we can find general expressions for the expected value and standard deviation of Y. We will look at the specific cases for $n = 1$, 2, and 3 and see what generalization this suggests. For the expected value, or mean, of the probability distribution, we have:

$n = 1$ $E(Y) = E(X_1) = 0q + 1p = p$
$n = 2$ $E(Y) = 0(q^2) + 1(2pq) + 2(p^2) = 2p(q + p) = 2p$

5. Using the distribution developed in Problem 4, show that $E(Y) = 3p$ when $n = 3$.

The obvious guess for a general result is that $E(Y) = np$ for distributions of this type, and that is correct. Thus, in a random sample of four adults from the city being studied, we would expect to see about $4(0.8) = 3.2$ high-school graduates.

6. The expected number of high-school graduates is not an integer. Provide a meaningful interpretation of this number.

When developing expressions for standard deviations, it is easier to begin with the variance, the square of the standard deviation, and then take a square root at the end. We denote variance by V, and recall that the variance is the expected value

of the square of the deviations from the expected value. Using the probability distribution for X_1, which has expected value p, the variance becomes

$$V(X_1) = E(X_1 - p)^2 = (0 - p)^2 q + (1 - p)^2 p = p^2 q + q^2 p$$
$$= pq(p + q) = pq$$

and the standard deviation is

$$SD(X_1) = \sqrt{V(X_1)} = \sqrt{pq}$$

7. For $n = 2$, show that $SD(Y) = \sqrt{2pq}$.

The above results suggest the correct generalization for the standard deviation, which is

$$SD(Y) = \sqrt{npq}$$

8. Find the standard deviation you would expect to see among the number of high-school graduates observed in samples of size 4. Repeat for samples of size 10, 20, and 40. What do you notice about the standard deviations for the counts Y that differs from the standard deviations for proportions studied in Lesson 7?

Look back at the probability distributions for $n = 1$ and $n = 2$, $n = 3$, the one developed in Problem 4. You can begin to generalize to a formula that works for any sample size and see why the distribution is called the ***binomial distribution***. The sum of the probabilities across all possible values in a probability distribution must equal 1. Exploring this idea for the various sample sizes leads to

$n = 1$ $\quad q + p = 1$
$n = 2$ $\quad q^2 + 2pq + p^2 = (q + p)^2 = 1$
$n = 3$ $\quad q^3 + 3pq^2 + 3p^2q + p^3 = (q + p)^3 = 1$

9. For the case $n = 2$, expand the binomial expansion $(q + p)^2$ to show that it equals the expression on the left. Do the same for the case $n = 3$.

The interesting result is that all of the probabilities for each value of n are terms of the binomial expansion of $(q + p)^n$. That is the reason for the name binomial probability distribution. It provides a way of writing a general probability statement for any sample size.

For a random sample of size n from a population with probability of "success" p on each selection, the probability that the

number of "successes" in the sample Y equals a specific count c is given by

$$P(Y = c) = \binom{n}{c} p^c (1-p)^{n-c}$$

where $\binom{n}{c}$ is the binomial coefficient calculated as $\binom{n}{c} = \frac{n!}{c!(n-c)!}$.

In this calculation, $n! = n(n-1)(n-2)\ldots 1$. This expression is called "n factorial."

10. Show that the formula for $P(Y = c)$ gives the expressions shown on the tables provided earlier in this lesson for the cases $n = 2$ and $n = 3$.

Practice and Applications

11. Here is a statement from the first page of this lesson.

It is reported that about 80% of adults over 25 in the U.S. are high-school graduates. Suppose a Gallup poll randomly selects only FOUR adults from a small city for which this population percent is thought to hold.

a. Write out the probability distribution for Y, the number of high-school graduates to be seen in a random sample of size 4. Use your calculator to find the numerical values.

b. Construct a bar graph of the probability distribution found in part a.

c. Find the expected value of Y. Mark the expected value of Y on the bar graph.

d. Find the standard deviation of Y.

e. Mark the point $E(Y)$ plus one standard deviation of Y on the bar graph. Similarly, mark the point $E(Y)$ minus one standard deviation of Y. What possible values of Y lie between these two points? What is the total probability for the values of Y between these two points?

12. A student is guessing at all five true-false questions on a quiz.

a. Give the probability distribution for the number of true-false questions the student gets correct.

b. If the student needs to answer at least 60% of the true-false questions in order to pass the quiz, what is the probability that the student will pass?

13. You are to take a true-false quiz for which you have not studied, so you must guess at all of the answers. You need to have at least 60% of the answers correct in order to pass. Would you rather have a 5-question quiz or a 10-question quiz? Explain your reasoning.

14. A blood bank knows that only about 10% of its regular donors have type-B blood.

a. Ten donors will appear at the blood bank today. What is the chance that the blood bank will get at least one donor with type-B blood?

b. One hundred blood donors will appear at the blood bank this month. What is the approximate probability that at least 10% of them will have type-B blood? Explain how you made this approximation.

c. The blood bank needs 16 type-B donors this month. If 100 donors appear this month, does the blood bank have a good chance of getting the amount of type-B blood it needs? What recommendation would you have for the blood-bank managers?

15. If Y represents the number of successes in a binomial distribution of n trials, then Y/n represents the proportion of successes in the n trials. Show that

$$E\left(\frac{Y}{n}\right) = p$$

and

$$SD\left(\frac{Y}{n}\right) = \sqrt{\frac{p(1-p)}{n}}$$

Is this consistent with what you learned in Lesson 7?

16. The median annual household income for U.S. households is about $39,000.

a. Among five randomly selected households, find the probability that four or more have incomes exceeding $39,000 per year.

b. Consider a random sample of 16 households.

i. What is the expected number of households with annual income below $39,000?

ii. What is the standard deviation of the number of households with annual income below $39,000?

iii. What is the approximate probability of seeing at least 10 of the 16 households with income below $39,000 annually? Show two different methods of answering this question.

c. Suppose in a sample of 16 households none had an annual income below $39,000. What might you suspect about this sample?

17. A home alarm system has detectors covering n zones of the house. Suppose the probability is 0.7 that a detector sounds an alarm when an intruder passes through its zone, and this probability is the same for all detectors. The alarms operate independently. An intruder enters the house and passes through all the zones.

a. What is the probability that an alarm sounds if $n = 3$?

b. What is the probability that an alarm sounds if $n = 6$? Is the probability of part a doubled?

c. Suppose the home owner wants the probability that an alarm will sound to be about 0.99. How large must n be?

18. From a large lot of new video tapes, n are to be sampled by a potential buyer and the number of defective tapes Y is to be observed. If at least one defective tape is seen in the sample of n tapes, the potential buyer will reject the entire lot. Find n so that the probability of detecting at least one defective tape is 0.5 in each case.

a. 10% of the lot is defective.

b. 5% of the lot is defective.

19. An oil-exploration firm is to drill 10 wells. Each well has a probability of 0.1 of producing oil. It costs the firm $50,000 to drill each well. A successful well will bring in oil worth $1,000,000.

a. Find the firm's expected gain from the 10 wells.

b. Find the standard deviation of the firm's gain for the 10 wells.

c. Is there a good chance that the firm will lose money on the 10 wells? A rough approximation is acceptable here.

d. Is there a good chance that the firm will gain more than $1.5 million from the 10 wells?

20. Ten CD players of a model no longer made are to be sold

for $100 each with a "double your money back" guarantee if the CD player fails in the first month of use. Suppose the probability of such a failure is 0.08. What is the expected net gain for the seller after all 10 CD players are sold? Ignore the original cost of the CD players to the seller.

SUMMARY

For a random sample of size n from a population with probability of "success" p on each selection, the probability that the number of "successes" in the sample Y equals a specific count c is given by

$$P(Y = c) = \binom{n}{c} p^c (1-p)^{n-c}$$

where $\binom{n}{c}$ is the binomial coefficient given by

$$\binom{n}{c} = \frac{n!}{c!(n-c)}$$

The expected value of Y and the standard deviation of Y are given, respectively, by

$$E(Y) = np$$

and

$$SD(Y) = \sqrt{npq}.$$

The random variable Y is said to have a ***binomial distribution.***

LESSON 9

The Geometric Distribution

What is the probability that your team will have to play four games before it gets its first loss?

What is the general shape of waiting-time distributions?

How can a mathematical model be developed for waiting-time distributions?

OBJECTIVES

Understand the basic properties of the geometric distribution.

Use the geometric distribution as a model for certain types of counts.

The opening example of Lesson 8 considered the problem of how many high-school graduates should appear in a random sample of four adults from a city. Suppose, however, that the interviewer is interested in finding only one high-school graduate from this city. A question of interest in this case might be, "How many people must be selected until a high-school graduate is found?" In other words, "How long will we have to wait to find the first success?" The first person selected might be a high-school graduate but, if not, a second person will have to be interviewed, and so on.

INVESTIGATE

How Long Must We Wait?

Describe another situation in which waiting time until "success" is an important consideration.

Discussion and Practice

1. Recall that the proportion of high-school graduates in the population is 0.80.

a. Can the number of adults that must be selected until the first high-school graduate is found be determined in advance of the sampling?

b. Could the number of adults selected until the first high-school graduate is found be quite large?

c. Is there a high probability that the number of adults selected until the first high-school graduate is found will be quite large? Explain your reasoning.

We will begin our investigation of the problem of waiting for a success by constructing a simulation for the situation presented above and then moving on to other cases. The simulations can be completed most efficiently if you work in pairs.

2. If you select adults one at a time, how many must be selected until the first high-school graduate is found? Recall that the proportion of high-school graduates in the population is 0.80. The number of the trial on which the first success occurs is a random variable, which we will denote by Y. Its distribution can be approximated by the following simulation.

   a. What is the probability that a high-school graduate is seen on the first selection? Find a device that will generate an event with this probability. You may use random-number tables, a random-number generator in your calculator, or some other device.

   b. Simulate the selection of the first adult by generating an event with the device selected in part a. Did you see a high-school graduate? In other words, was your waiting time to success just one interview?

   c. If you had success on the first selection, then stop this run of the simulation. If you did not have success on the first trial, then continue generating events until the first success occurs. Record a value for Y, the number of the trial on which the first success occurred. NOTE: This number must be at least 1.

   d. Repeat parts b and c ten times, recording a value for Y each time.

   e. Combine your values of Y with those of your classmates.

i. Construct a plot to represent the simulated distribution of Y.

ii. Describe the shape of this distribution.

iii. Approximate the mean of this distribution.

3. Each fly of a certain species has a 0.5 chance of dying in any one-day period after its birth. Simulate the probability distribution of Y, the age of flies of this species. In this simulation, "age" can be thought of as the number of days until death, or the waiting time until death. Instead of generating the values of Y one at a time, as we did in Problem 2, consider a simulation that gives a whole set of Y values simultaneously.

a. Gather a set of about 200 pennies. Place the pennies in a container, shake well, and toss them onto a table. Count the heads.

b. If "head on a coin" represents "death of a fly" what fraction of flies died during the first day? Does this seem like a reasonable estimate of the probability that a fly died during the first day, $Y = 1$? Explain.

c. Gather the pennies that came up tails on the first toss. Place them in the container, shake well, and toss them onto a table. Count the heads. What fraction of the *original number* of pennies showed the first head on the second toss? What is the approximate probability that a fly dies during the second day, $Y = 2$?

d. Continue gathering the tails, tossing them, and counting the heads until there are five or fewer tails left.

i. Record the values of Y and the approximate probability for each.

ii. Show the simulated probability distribution for Y on an appropriate graph.

iii. Describe the shape of this distribution.

iv. Approximate the mean of this distribution.

4. Compare the methods of simulation used in Problems 2 and 3. What would be the difficulty of using the method of Problem 3 on Problem 2?

5. To provide more experience with probability distributions of waiting times, we have simulated the cases for $p = 0.3$, $p = 0.5$, and $p = 0.8$, where p is the probability of success

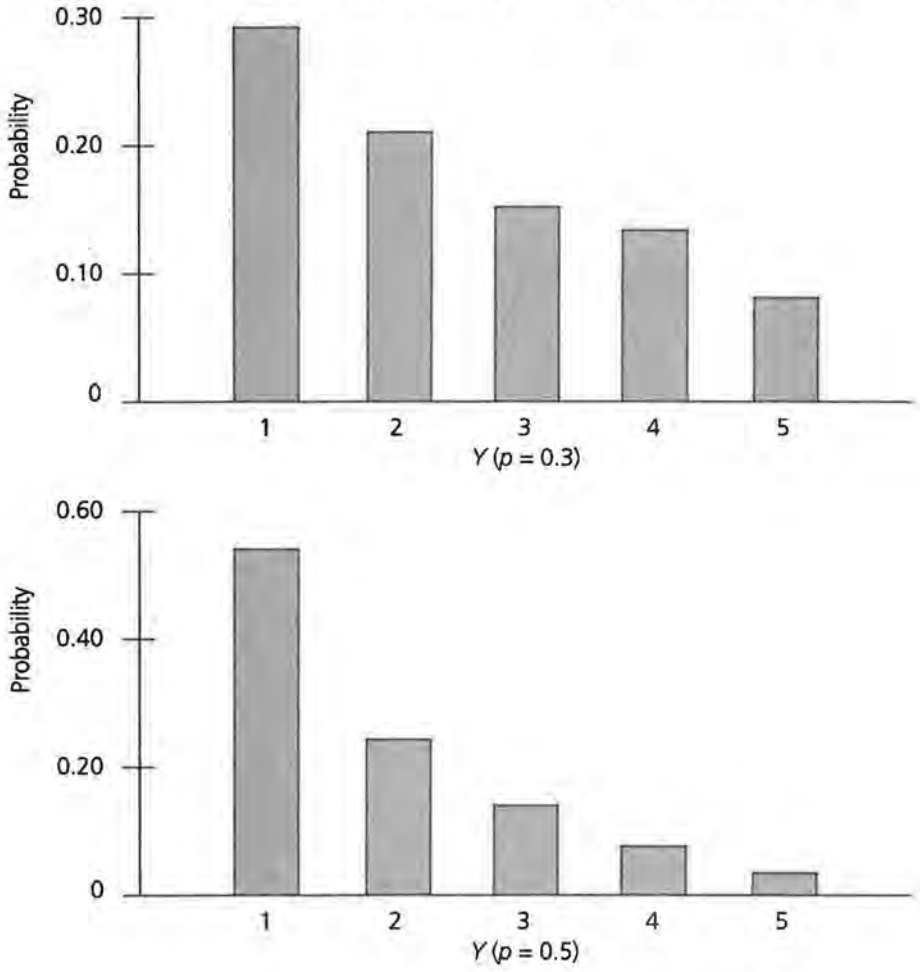
on any one event. These are shown below. The latter two cases should have similarities to those generated in Problems 2 and 3.

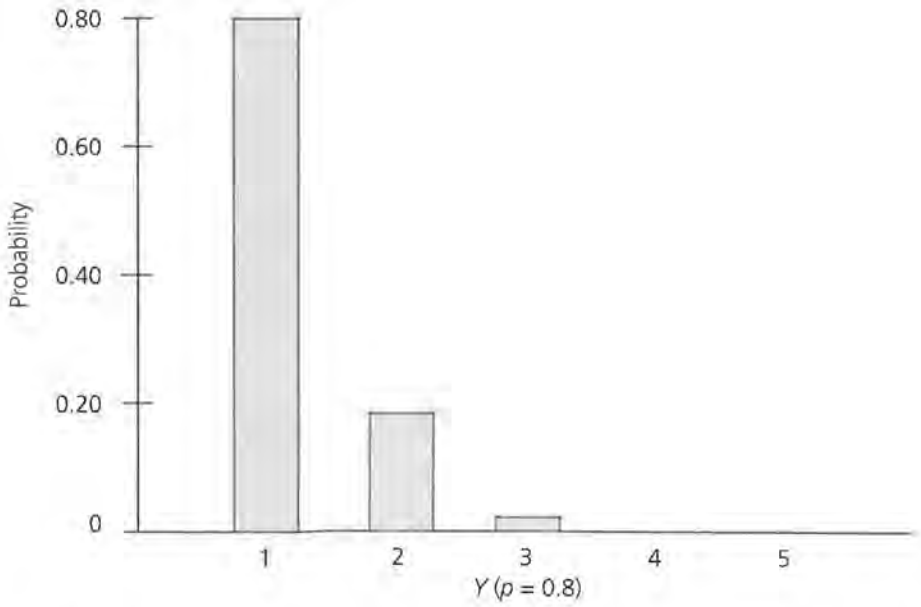
a. How are the shapes of the three distributions similar? How are they different?

b. Do the probabilities shown in each distribution appear to add to 1? If not, why not?

c. Approximate the mean of each distribution.

Simulated Waiting-Timing Distributions





The Geometric Distribution

Simulation provides a good way to investigate properties of distributions, but it is inefficient to run a simulation every time a new problem involving one of these waiting-time distributions arises. It turns out that a formula for the probability distribution of Y can be derived quite easily—more easily, in fact, than in the binomial case of Lesson 8.

Suppose we are sampling people from a large population in which the proportion of people having the characteristic labeled success is p. Using notation similar to that of Lesson 8, we can record the outcome of each selection in sequence by letting

$X_i = 1$ if the ith selection in the sequence is a success

and

$X_i = 0$ if the ith selection in the sequence is not a success.

Then, $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p = q$.

6. Development of a formula for the probability distribution of Y, the selection on which the first success occurs, proceeds along the following lines.

a. Write $P(Y = 2)$ in terms of the random variables X_i. Do the same for $P(Y = 3)$.

b. Write $P(Y = c)$ in terms of the random variables X_i, for any positive integer c.

c. Using the result of part b, write $P(Y = c)$ in terms of p and q. Make the result as compact as possible.

d. What assumption is necessary for the probability calculations used in part c?

With the answer to Problem 6, you can now see what the probability distribution for the random variable Y looks like in table form.

Y	$P(Y)$
1	p
2	qp
3	q^2p
4	q^3p
.	.
.	.
.	.

The sum of these probabilities form an infinite series, and the sum of the probabilities across all possible values of Y form an infinte sum:

$$\sum_{c=1}^{\infty} q^{c-1}p = p\sum_{c=1}^{\infty} q^{c-1} = p(1 + q + q^2 + q^3 + q^4 + \ldots) = p\left(\frac{1}{1-q}\right) = 1.$$

The infinite sum inside the brackets is called a ***geometric progression***. The basic result for summing such a series is as follows:

$$a + ax + ax^2 + ax^3 + \ldots = \frac{a}{1-x}$$

as long as the absolute value of x is less than 1.

You see, now, why the probability distribution developed in this lesson is called the ***geometric distribution.***

7. The expected value of the geometric distribution can be found by summing an array of infinite series. Show that $E(Y)$, for the geometric random variable Y, can be written in a triangular array as

$$\begin{aligned} E(Y) = p(1 &+ q + q^2 + q^3 + \ldots \\ &+ q + q^2 + q^3 + \ldots \\ &+ q^2 + q^3 + \ldots \\ &+ q^3 + \ldots \\ &+ \ldots) \end{aligned}$$

a. Sum each row of the array inside the brackets by using the result for geometric progressions.

b. Find $E(Y)$ by summing the row totals you just found, again using the result for geometric progressions.

c. Does your result for $E(Y)$ make sense from a practical point of view? Explain.

8. The standard deviation of a geometric random variable Y is difficult to find from the basic distribution, but it turns out to be $SD(Y) = \frac{1}{p}\sqrt{1-p}$.

a. Find the expected value and standard deviation for each of the three geometric distributions simulated on pages 79–80.

b. Plot the expected value and points one standard deviation above and below the expected value on each distribution on pages 79–80.

c. Does the standard deviation appear to be a good measure of a typical deviation from the mean here? Explain.

Practice and Applications

9. In the scenario outlined at the beginning of this lesson, the probability of randomly selecting a high-school graduate was 0.8.

a. What is the probability of getting the first high-school graduate on the third selection?

b. What is the probability that it will take at least three selections to get the first high-school graduate?

c. Suppose the first high-school graduate comes up on the third selection. What is the probability that it will take at least three more selections to obtain a second high-school graduate?

d. What is the expected number of selections to obtain the first high-school graduate?

e. What is the expected number of selections to obtain two high-school graduates?

10. Suppose 10% of the engines manufactured on a certain assembly line have at least one defect. Engines are randomly sampled from this line one at a time and tested.

a. What is the probability that the first non-defective engine is found on the third trial?

b. What is the expected number of engines that need to be

tested before the first non-defective engine is found?

c. What is the standard deviation of the number of engines that need to be tested before the first non-defective engine is found?

d. Suppose it costs \$100 to test one engine. What are the expected value and the standard deviation of the cost of inspection up to and including the first non-defective engine? Will the cost of inspection often exceed \$200? Explain.

11. The telephone lines coming into an airline reservation office are all busy about 60% of the time.

a. If you are calling this office, what is the probability that it will take you only one try to get through? Two tries? Four tries?

b. What is your expected number of tries to complete the call?

12. An oil-exploration firm is to drill wells at a particular site until it finds one that will produce oil. Each well has a probability of 0.1 of producing oil. It costs the firm \$50,000 to drill each well.

a. What is the expected number of wells to be drilled?

b. What are the expected value and the standard deviation of the cost of drilling to get the first successful well?

c. What is the probability that it will take at least five tries to get the first successful well? At least 15?

SUMMARY

For a sequence of selections from a large population in which the probability of "success" p stays the same for all selections, the number of the selection on which the first success occurs Y has a ***geometric distribution***. The mean and the standard deviation of this distribution are given by

$$E(Y) = \frac{1}{p}$$

and

$$SD(Y) = \frac{1}{p}\sqrt{1-p}$$

Lessons 8 and 9

1. In the population as a whole, about 46% of people have type-O blood, about 42% have type-A, about 8% have type-B, and about 4% have type-AB.
   a. In a random sample of five people, how many would be expected to have type-A blood? How many would be expected to have either type-A or type-AB blood?
   b. For a random sample of five people, find the probability distribution for the number having type-AB blood. Describe the shape of this distribution.
   c. A total of 200 people are to donate blood at a certain blood bank this week. Find the expected value and the standard deviation of the number of type-A donors the blood bank will see this week.
   d. The blood bank in part c needs 100 donors of type-A blood this week. If 200 donors appear this week, is there a good chance that it will get the number it needs? Explain your reasoning.
   e. If the donors to a blood bank come in sequential order, what is the chance that more than four donors will have to be tested before the first one with type-A blood shows up? Answer the same question for type-AB blood.
   f. If the donors to a blood bank come in sequential order, what is the expected number of donors that must be tested in order to find the first one with type-AB blood?
2. Refer to the percents of blood types given in Problem 1. The Rh, or *Rhesus*, factor in the blood is independent of the blood type. About 85% of people are Rh positive.
   a. For a random sample of five donors, find the probability distribution of the number of type O-negative donors that will be seen. Describe the shape of this distribution.
   b. A total of 200 people are to donate blood in a certain blood bank this week. Find the expected value and the standard deviation of the number of A-positive donors.

If the blood bank needs 80 A-positive donors this week, does it stand a good chance of getting them? Explain.

c. The blood bank is in need of an A-positive donor as soon as possible. Describe the distribution of the number of donors that must be tested to find the first A-positive donor. What is the expected number of donors to be tested in order to find one who is A positive?

3. You are to take a ten-question multiple choice exam. Each question has four choices, of which only one is correct. You know none of the answers and decide to guess at an answer for each question.

a. Describe how to set up a simulation for the probability distribution of the number of questions answered correctly.

b. What is the probability that you will answer at least 60% of the questions correctly?

c. Suppose the teacher changes the questions so that each one has three choices, one of which is correct. Would your chance of answering at least 60% correctly by guessing go up or down from the answer on Problem 3b? Explain.

d. Refer to the original scenario of four choices per question. Suppose the test now has five questions rather than ten. Does your chance of getting at least 60% correct by guessing increase or decrease from your chance in Problem 3b? Explain.

4. Each box of a certain brand of cereal contains a coupon that can be redeemed for a poster of a famous sports figure. There are five different coupons, representing five different sports figures.

a. Suppose you are interested in one sports figure in particular. Set up a simulation that would produce an approximate probability distribution for the number of boxes of cereal you would have to buy in order to get one coupon for that particular poster. You stop buying cereal when you get the poster you want.

b. What is the probability that you would get the coupon for the particular sports figure of interest in four or fewer boxes of cereal? You stop buying cereal when you get the poster you want.

c. What is the expected number of boxes of cereal you would have to buy to get the one coupon you want?

d. Suppose you want to get two particular posters out of the five available. What is the expected number of boxes of cereal you would have to buy to get the two specific ones? (HINT: At the outset, the probability of getting a coupon you want in any one box is $\frac{2}{5}$. After you get one of them, what is the probability of getting the other coupon you want? Think of the expected number of boxes of cereal being purchased in terms of these two stages.)

Dale Seymour Publications® is a leading publisher of K–12 educational materials in mathematics, thinking skills, science, language arts, and art education.

DALE SEYMOUR PUBLICATIONS®

ISBN 1-57232-240-3
21179