7.4 Hypergeometric Distributions

When choosing the starting line-up for a game, a coach obviously has to choose a different player for each position. Similarly, when a union elects delegates for a convention or you deal cards from a standard deck, there can be no repetitions. In such situations, each selection reduces the number of items that could be selected in the next trial. Thus, the probabilities in these trials are dependent. Often we need to calculate the probability of a specific number of successes in a given number of dependent trials.

INVESTIGATE & INQUIRE: Choosing a Jury

In Ontario, a citizen can be called for jury duty every three years. Although most juries have 12 members, those for civil trials in Ontario usually require only 6 members. Suppose a civil-court jury is being selected from a pool of 18 citizens, 8 of whom are men. Develop a simulation to determine the probability distribution for the number of women selected for this jury.

1. Select a random-number generator to simulate the selection process.
2. Decide how to simplify the selection process. Decide, also, whether the full situation needs to be simulated or whether a proportion of the trials would be sufficient.
3. Design each trial so that it simulates the actual situation. Ensure that each trial is dependent by setting the random-number generator so that there are no repetitions within each series of trials.
4. Set up a method to record the number of successes in each experiment. Pool your results with those of other students in your class, if necessary.
5. Use the results to estimate the probabilities of x successes (women) in r trials (selections of a juror).
6. Reflect on the results. Do they accurately represent the probability of x women being selected?
7. Compare your simulation and its results with those of your classmates. Which are the better simulations? Explain why.

Data in Action

The cost of running the criminal, civil, and family courts in Ontario was about $310 million for 2001. These courts have the equivalent of 3300 full-time employees.

The simulation in the investigation models a **hypergeometric distribution**. Such distributions involve a series of *dependent* trials, each with success or failure as the only possible outcomes. The probability of success changes as each trial is made. The random variable is the number of successful trials in an experiment. Calculations of probabilities in a hypergeometric distribution generally require formulas using combinations.

Example 1 Jury Selection

a) Determine the probability distribution for the number of women on a civil-court jury selected from a pool of 8 men and 10 women.

b) What is the expected number of women on the jury?

Solution 1 Using Pencil and Paper

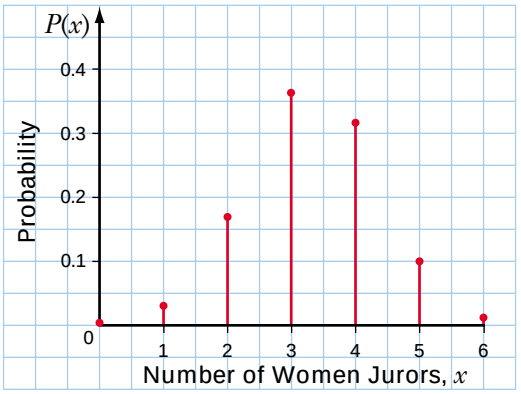
a) The selection process involves dependent events since each person who is already chosen for the jury cannot be selected again. The total number of ways the 6 jurors can be selected from the pool of 18 is

$$n(S) = {}_{18}C_6$$
$$= 18\ 564$$

This combination could also be written as C(18, 6) or $\binom{18}{6}$.

There can be from 0 to 6 women on the jury. The number of ways in which x women can be selected is ${}_{10}C_x$. The men can fill the remaining $6 - x$ positions on the jury in ${}_8C_{6-x}$ ways. Thus, the number of ways of selecting a jury with x women on it is ${}_{10}C_x \times {}_8C_{6-x}$ and the probability of a jury with x women is

$$P(x) = \frac{n(x)}{n(S)}$$
$$= \frac{{}_{10}C_x \times {}_8C_{6-x}}{{}_{18}C_6}$$



Number of Women, x	Probability, $P(x)$
0	$\frac{{}_{10}C_0 \times {}_8C_6}{{}_{18}C_6} \doteq 0.001\ 51$
1	$\frac{{}_{10}C_1 \times {}_8C_5}{{}_{18}C_6} \doteq 0.030\ 17$
2	$\frac{{}_{10}C_2 \times {}_8C_4}{{}_{18}C_6} \doteq 0.169\ 68$
3	$\frac{{}_{10}C_3 \times {}_8C_3}{{}_{18}C_6} \doteq 0.361\ 99$
4	$\frac{{}_{10}C_4 \times {}_8C_2}{{}_{18}C_6} \doteq 0.316\ 74$
5	$\frac{{}_{10}C_5 \times {}_8C_1}{{}_{18}C_6} \doteq 0.108\ 60$
6	$\frac{{}_{10}C_6 \times {}_8C_0}{{}_{18}C_6} \doteq 0.011\ 31$

b) $E(X) = \sum_{i=0}^{6} x_i P(x_i)$

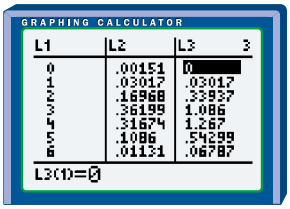
$\doteq (0)(0.001\ 51) + (1)(0.030\ 17) + (2)(0.169\ 68) + (3)(0.361\ 99)$
$+ (4)(0.316\ 74) + (5)(0.108\ 60) + (6)(0.011\ 31)$
$\doteq 3.333\ 33$

The expected number of women on the jury is approximately 3.333.

Solution 2 Using a Graphing Calculator

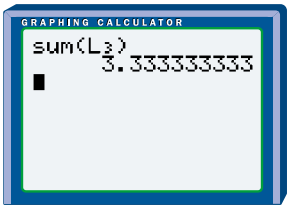
a) Enter the possible values for x, 0 to 6, in L1. Then, enter the formula for $P(x)$ in L2:

(10 nCr L1) × (8 nCr (6–L1)) ÷ (18 nCr 6)



b) Calculate $xP(x)$ in L3 using the formula L1 × L2.

QUIT to the home screen. You can find the expected number of women by using the sum(function in the LIST MATH menu.



The expected number of women on the jury is approximately 3.333.

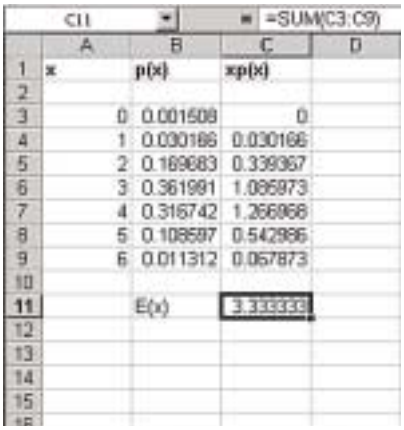
Solution 3 Using a Spreadsheet

a) Open a new spreadsheet. Create titles x, p(x), and xp(x) in columns A to C.

Enter the values of the random variable x in column A, ranging from 0 to 6. Next, use the combinations function to enter the formula for $P(x)$ in cell B3 and copy it to cells B4 through B9.

b) Calculate $xP(x)$ in column C by entering the formula A3*B3 in cell C3 and copying it to cells C4 through C9. Then, calculate the expected value using the **SUM function**.

The expected number of women on the jury is approximately 3.333.



	A	B	C	D
1	x	p(x)	xp(x)	
2				
3	0	0.001508	0	
4	1	0.030166	0.030166	
5	2	0.169683	0.339367	
6	3	0.361991	1.085973	
7	4	0.316742	1.266966	
8	5	0.108597	0.542986	
9	6	0.011312	0.067873	
10				
11		E(x)	3.333333	
12				
13				
14				
15				

Solution 4 Using Fathom™

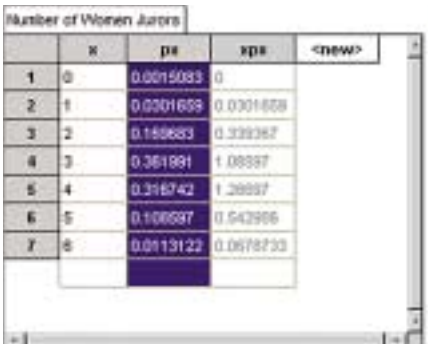
Open a new Fathom™ document. Drag a new **collection** box to the work area and name it Number of Women Jurors. Create seven new cases.

Drag a new **case table** to the work area. Create three new attributes: x, px, and xpx. Enter the values from 0 to 6 for the x attribute. Right-click on the px attribute, select Edit Formula, and enter

combinations(10,x)*combinations(8,6-x)/combinations(18,6)

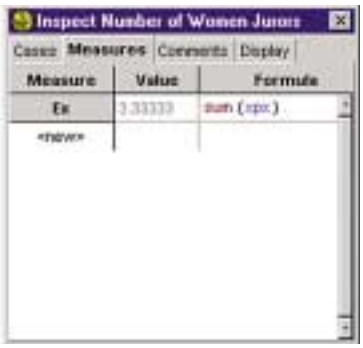
Similarly, calculate $xP(x)$ using the formula x*px. Next, double-click on the **collection** box to open the **inspector**. Select the Measures tab, and name a new measure Ex. Right-click on Ex and use the **sum function** to enter the formula sum(x*px).

The expected number of women on the jury is approximately 3.333.



Number of Women Jurors

	x	px	xpx	<new>
1	0	0.0015083	0	
2	1	0.0301659	0.0301659	
3	2	0.169683	0.339367	
4	3	0.361991	1.08597	
5	4	0.316742	1.26697	
6	5	0.108597	0.542986	
7	6	0.0113122	0.0678733	



Inspect Number of Women Jurors

Cases | Measures | Comments | Display

Measure	Value	Formula
Ex	3.33333	sum (xpx)
<new>		

You can generalize the methods in Example 1 to show that for a hypergeometric distribution, the probability of x successes in r dependent trials is

Probability in a Hypergeometric Distribution

$$P(x) = \frac{{}_aC_x \times {}_{n-a}C_{r-x}}{{}_nC_r},$$

where a is the number of successful outcomes among a total of n possible outcomes.

Although the trials are dependent, you would expect the *average* probability of a success to be the same as the ratio of successes in the population, $\frac{a}{n}$. Thus, the expectation for r trials would be

Expectation for a Hypergeometric Distribution

$$E(X) = \frac{ra}{n}$$

This formula can be proven more rigorously by some challenging algebraic manipulation of the terms when $P(x) = \frac{{}_aC_x \times {}_{n-a}C_{r-x}}{{}_nC_r}$ is substituted into the equation for the expectation of any probability distribution, $E(X) = \sum_{i=1}^{n} x_i P(x_i)$.

Example 2 Applying the Expectation Formula

Calculate the expected number of women on the jury in Example 1.

Solution

$$\begin{aligned} E(X) &= \frac{ra}{n} \\ &= \frac{6 \times 10}{18} \\ &= 3.3\overline{3} \end{aligned}$$

The expected number of women jurors is $3.3\overline{3}$.

Example 3 Expectation of a Hypergeometric Distribution

A box contains seven yellow, three green, five purple, and six red candies jumbled together.

a) What is the expected number of red candies among five candies poured from the box?

b) Verify that the expectation formula for a hypergeometric distribution gives the same result as the general equation for the expectation of any probability distribution.

Solution

a) $n = 7 + 3 + 5 + 6 \qquad r = 5 \qquad a = 6$
$= 21$

Using the expectation formula for the hypergeometric distribution,

$$E(X) = \frac{ra}{n}$$
$$= \frac{5 \times 6}{21}$$
$$= 1.4285\ldots$$

One would expect to have approximately 1.4 red candies among the 5 candies.

b) Using the general formula for expectation,

$$E(X) = \sum xP(x)$$
$$= (0)\frac{{}_6C_0 \times {}_{15}C_5}{{}_{21}C_5} + (1)\frac{{}_6C_1 \times {}_{15}C_4}{{}_{21}C_5} + (2)\frac{{}_6C_2 \times {}_{15}C_3}{{}_{21}C_5} + (3)\frac{{}_6C_3 \times {}_{15}C_2}{{}_{21}C_5} + (4)\frac{{}_6C_4 \times {}_{15}C_1}{{}_{21}C_5} + (5)\frac{{}_6C_5 \times {}_{15}C_0}{{}_{21}C_5}$$
$$= 1.4285\ldots$$

Again, the expected number of red candies is approximately 1.4.

Example 4 Wildlife Management

In the spring, the Ministry of the Environment caught and tagged 500 raccoons in a wilderness area. The raccoons were released after being vaccinated against rabies. To estimate the raccoon population in the area, the ministry caught 40 raccoons during the summer. Of these 15 had tags.

a) Determine whether this situation can be modelled with a hypergeometric distribution.

b) Estimate the raccoon population in the wilderness area.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

To learn more about sampling and wildlife, visit the above web site and follow the links. Write a brief description of some of the sampling techniques that are used.

Solution

a) The 40 raccoons captured during the summer were all different from each other. In other words, there were no repetitions, so the trials were dependent. The raccoons were either tagged (a success) or not (a failure). Thus, the situation does have all the characteristics of a hypergeometric distribution.

b) Assume that the number of tagged raccoons caught during the summer is equal to the expectation for the hypergeometric distribution. You can substitute the known values in the expectation formula and then solve for the population size, n.

Here, the number of raccoons caught during the summer is the number of trials, so $r = 40$. The number of tagged raccoons is the number of successes in the population, so $a = 500$.

$$E(X) = \frac{ra}{n}, \quad \text{so} \quad 15 \doteq \frac{40 \times 500}{n}$$
$$n \doteq \frac{40 \times 500}{15}$$
$$n \doteq 1333.3$$

The raccoon population in the wilderness area is approximately 1333.

Alternatively, you could assume that the proportion of tagged raccoons among the sample captured during the summer corresponds to that in the whole population. Then, $\frac{15}{40} = \frac{500}{n}$, which gives the same estimate for n as the calculation shown above.

Key Concepts

- A hypergeometric distribution has a specified number of dependent trials having two possible outcomes, success or failure. The random variable is the number of successful outcomes in the specified number of trials. The individual outcomes cannot be repeated within these trials.
- The probability of x successes in r dependent trials is $P(x) = \frac{{}_aC_x \times {}_{n-a}C_{r-x}}{{}_nC_r}$, where n is the population size and a is the number of successes in the population.
- The expectation for a hypergeometric distribution is $E(X) = \frac{ra}{n}$.
- To simulate a hypergeometric experiment, ensure that the number of trials is representative of the situation and that each trial is dependent (no replacement or resetting between trials). Record the number of successes and summarize the results by calculating probabilities and expectation.

Communicate Your Understanding

1. Describe how the graph in Example 1 differs from the graphs of the uniform, binomial, and geometric distributions.
2. Consider this question: What is the probability that 5 people out of a group of 20 are left handed if 10% of the population is left-handed? Explain why this situation does not fit a hypergeometric model. Rewrite the question so that you can use a hypergeometric distribution.

Practise

A

1. Which of these random variables have a hypergeometric distribution? Explain why.

a) the number of clubs dealt from a deck

b) the number of attempts before rolling a six with a die

c) the number of 3s produced by a random-number generator

d) the number of defective screws in a random sample of 20 taken from a production line that has a 2% defect rate

e) the number of male names on a page selected at random from a telephone book

f) the number of left-handed people in a group selected from the general population

g) the number of left-handed people selected from a group comprised equally of left-handed and right-handed people

2. Prepare a table and a graph of a hypergeometric distribution with

a) $n = 6, r = 3, a = 3$ **b)** $n = 8, r = 3, a = 5$

Apply, Solve, Communicate

B

3. There are five cats and seven dogs in a pet shop. Four pets are chosen at random for a visit to a children's hospital.

a) What is the probability that exactly two of the pets will be dogs?

b) What is the expected number of dogs chosen?

4. **Communication** Earlier this year, 520 seals were caught and tagged. On a recent survey, 30 out of 125 seals had been tagged.

a) Estimate the size of the seal population.

b) Explain why you cannot calculate the exact size of the seal population.

5. Of the 60 grade-12 students at a school, 45 are taking English. Suppose that 8 grade-12 students are selected at random for a survey.

a) Develop a simulation to determine the probability that 5 of the selected students are studying English.

b) Use the formulas developed in this section to verify your simulation results.

6. **Inquiry/Problem Solving** In a study of Canada geese, 200 of a known population of 1200 geese were caught and tagged. Later, another 50 geese were caught.

a) Develop a simulation to determine the expected number of tagged geese in the second sample.

b) Use the formulas developed in this section to verify your simulation results.

7. **Application** In a mathematics class of 20 students, 5 are bilingual. If the class is randomly divided into 4 project teams,

a) what is the probability that a team has fewer than 2 bilingual students?

b) what is the expected number of bilingual students on a team?

8. In a swim meet, there are 16 competitors, 5 of whom are from the Eastern Swim Club.

a) What is the probability that 2 of the 5 swimmers in the first heat are from the Eastern Swim Club?

b) What is the expected number of Eastern Swim Club members in the first heat?

9. The door prizes at a dance are four \$10 gift certificates, five \$20 gift certificates, and three \$50 gift certificates. The prize envelopes are mixed together in a bag, and five prizes are drawn at random.

a) What is the probability that none of the prizes is a \$10 gift certificate?

b) What is the expected number of \$20 gift certificates drawn?

10. A 12-member jury for a criminal case will be selected from a pool of 14 men and 11 women.

a) What is the probability that the jury will have 6 men and 6 women?

b) What is the probability that at least 3 jurors will be women?

c) What is the expected number of women?

11. Seven cards are dealt from a standard deck.

a) What is the probability that three of the seven cards are hearts?

b) What is the expected number of hearts?

12. A bag contains two red, five black, and four green marbles. Four marbles are selected at random, without replacement. Calculate

a) the probability that all four are black

b) the probability that exactly two are green

c) the probability that exactly two are green and none are red

d) the expected numbers of red, black, and green marbles

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

13. A calculator manufacturer checks for defective products by testing 3 calculators out of every lot of 12. If a defective calculator is found, the lot is rejected.

a) Suppose 2 calculators in a lot are defective. Outline two ways of calculating the probability that the lot will be rejected. Calculate this probability.

b) The quality-control department wants to have at least a 30% chance of rejecting lots that contain only one defective calculator. Is testing 3 calculators in a lot of 12 sufficient? If not, how would you suggest they alter their quality-control techniques to achieve this standard? Support your answer with mathematical calculations.

C

14. Suppose you buy a lottery ticket for which you choose six different numbers between 1 and 40 inclusive. The order of the first five numbers is not important. The sixth number is a bonus number. To win first prize, all five regular numbers and the bonus number must match, respectively, the randomly generated winning numbers for the lottery. For the second prize, you must match the bonus number plus four of the regular numbers.

a) What is the probability of winning first prize?

b) What is the probability of winning second prize?

c) What is the probability of not winning a prize if your first three regular numbers match winning numbers?

15. **Inquiry / Problem Solving** Under what conditions would a binomial distribution be a good approximation for a hypergeometric distribution?

16. **Inquiry / Problem Solving** You start at a corner five blocks south and five blocks west of your friend. You walk north and east while your friend walks south and west at the same speed. What is the probability that the two of you will meet on your travels?

17. A research company has 50 employees, 20 of whom are over 40 years old. Of the 22 scientists on the staff, 12 are over 40. Compare the expected numbers of older and younger scientists in a randomly selected focus group of 10 employees.