Section IV: Probability

Investigation 12

Chances of Getting the Flu?

Simulations

Overview

This investigation develops a probability distribution through the design and use of a simulation. It follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are formulate a statistical question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level B activity.

This activity is based on a simulation problem from *The Art and Techniques of Simulation*, published by Dale Seymour and the American Statistical Association. (This module is part of the Quantitative Literacy Series. Though out of print, the book is available through book resale sites and on Amazon.com.)

Instructional Plan

Brief Overview

» Read and discuss the scenario about the spread of flu in an apartment building.

» Formulate the statistical/probabilistic question: "What is an estimate for the probability that all six people who live in an apartment building will get the flu?"

» Demonstrate the steps to conduct a simulation to answer the probabilistic question.

» Have students conduct the simulation using a die or technology and report their results.

» Collect class data in a table, convert the results to relative frequencies and a probability distribution.

» Use the probability distribution to answer the statistical/probabilistic question.

Hand out Student Worksheet 12.1 Flu Epidemic. Direct your students to read the first paragraph in the scenario.

Scenario

Did you get a flu vaccine last year? If so, did you still get the flu?

Infectious diseases (or diseases that are often caused by a bacteria or virus) are extensively researched in the medical field. These diseases result in colds, seasonal flu, and major epidemics that affect large numbers of people or animals in some cases.

In the fall of 1918, a flu pandemic erupted and became one of the greatest loss of lives the world had ever seen. By many accounts, the flu claimed between 2.5% and 5% of the global population. At that time, there was no flu vaccine, no antiviral drugs, and no antibiotics to help lessen the number of patients who got the flu or aid in the recovery from the flu.

As a result of this pandemic, countries began to put a greater emphasis on the study of patterns, causes, and effects of diseases.

Medical researchers are actively involved in understanding what causes the disease, how it is spread, how long it lasts, and other data related to the health of patients.

Source: www.smithsonianmag.com/history/how-1918-flu-pandemic-revolutionized-public-health-180965025

Learning Goals

» Design and carry out a simulation to estimate the probability of a random event.

» Develop a non-uniform probability distribution based on a simulation.

Mathematical Practices Through a Statistical Lens

MP5. Use appropriate tools strategically.

Statistically proficient students are able to use technological tools to carry out simulations for exploring and deepening their understanding of statistical and probabilistic concepts.

Materials

Student worksheets are available at *www.statisticsteacher.org/statistics-teacher-publications/focus*.

» Large foam die

» Die for each pair of students

» Technology like the TI-84 graphing calculator with ProbSim app or a similar rolling die application such as *www.random.org/dice*

» Student Worksheet 12.1 Flu Epidemic

» Student Worksheet 12.2 Simulation Steps

» Exit Ticket

Estimated Time

One 50-minute class period

Pre-Knowledge

Students should already be able to find the probability of simple events.

Students understand the probability of an event E is equal to:

$$P(E) = \frac{\text{Number of trials favorable to E}}{\text{Total number of trials in the experiment}}$$

Discuss with students the flu scenario and ask what type of precautions they can take to avoid getting the flu.

Ask your students to read the flu example.

Flu Example

Consider the following simple example of an infectious disease, like a cold or flu, and how it spreads throughout a small apartment building.

Suppose a strain of the flu has a one-day infection period (i.e., a person with the flu can only infect another person for one day and, after that day, the person can't spread the flu and is immune—that is, once you get the flu, you can't get this strain of flu again). This strain of flu is potent; if a person comes into contact with someone with the flu, that person will get the flu for certain.

Six people live in a small apartment building. One person catches this very infectious strain of flu and randomly encounters one of the other tenants during the infection period, and this second tenant gets this strain of flu. This second tenant infected with the flu visits a third tenant at random during the next day, and this third tenant gets the flu. The process continues with a newly infected person randomly visiting someone who hasn't had the flu or visiting an immune person and the strain of flu dies out. If an infected person visits an immune person, then the spread of the flu will end, as the flu in this example has only a one-day infection period.

Ask your students to summarize how this strain of flu spreads.

What is the least number of tenants who could get the flu?

Answer: *Two tenants, The first tenant gets the flu and visits a second tenant, who then goes back and visits the first tenant.*

What is the highest number of tenants who could get the flu?

Answer: *All six tenants*

Formulate a Statistical Question

Discuss with your students that one way to investigate an estimate of the number of people who would get the flu in this apartment building is to design and conduct a simulation. A simulation is a procedure developed for answering questions about real problems by running experiments that resemble the real-life situation. Instead of finding a large number of apartment buildings with six apartments and one person with the flu, a simulation could be designed to provide outcomes of the number of people who get the flu.

Ask students to consider the statistical/probabilistic question: "What is an estimate for the probability that all six people who live in an apartment building will get the flu?"

Collect Appropriate Data

To help your students understand the scenario, conduct a simulation involving them.

» Select six students and have them come to the front of the room. These six students represent the people living in the apartment building. Number each student from 1 to 6.

» *Day 1:* Roll the large foam die to determine Patient Zero, who will have the flu first. For example, if a 3 is rolled, then Person 3 has the flu. Have Person 3 roll the die, and then have Person 3 visit the person whose number is rolled. For example, a 4 is rolled. Remember this flu is potent; if a person is "visited," they will get the flu. Now two people have gotten the flu—persons 3 and 4. If Person 3 rolled a 3, then Person 3 would roll again since a person can't visit him/herself.

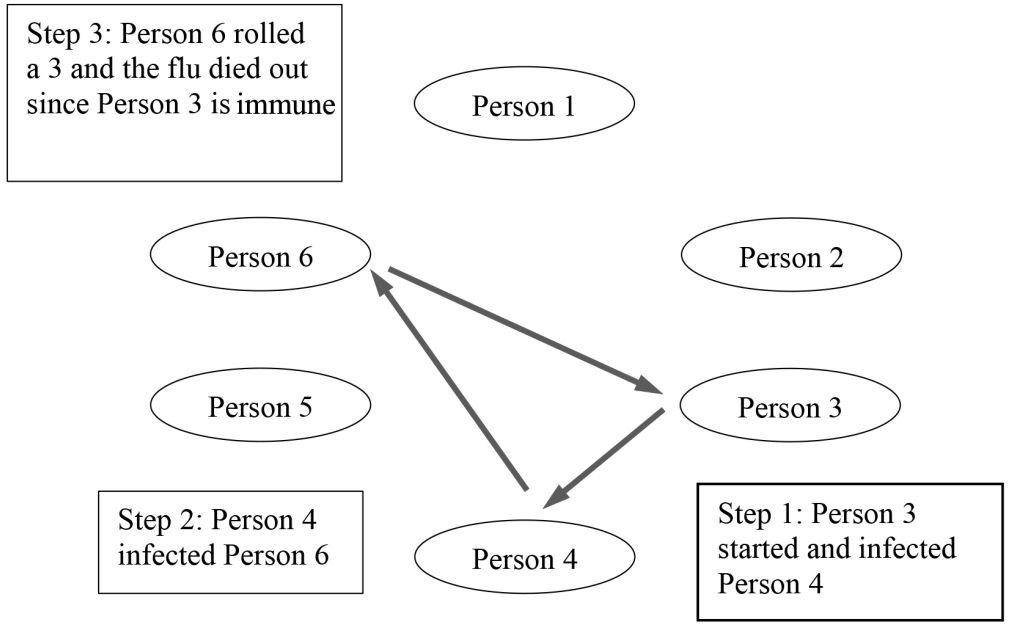

Figure 12.1

» *Day 2:* Person 3 is now immune (once you have had the flu, you can't get it again and you are no longer contagious) and Person 4, who now has the flu rolls a die and visits (infects) the person whose number was selected. For example, Person 6. Three people (3, 4, and 6) now have had the flu, unless Person 4 was to roll a 3. In that case, the flu would die out since the infected person visited a person who already had the flu. If the person rolls his/her own number, have the person roll again since a person can't visit him/herself.

» *Day 3:* Person 3 and Person 4 are immune. Person 6, who now has the flu, rolls a die and visits a person. Continue until a person visits someone who has already had the flu (i.e., immune) or someone who has not been infected. If the person rolls his or her own number, have the person roll again since a person can't visit him or herself.

Make a note of the number of people who got the flu.

Note: Students could also draw six circles—one for each person in the apartment building—and draw lines connecting the circles to show how the flu spreads as the simulation progresses.

Figure 12.1 illustrates the example above showing one trial in which three people were infected before the flu died out.

Emphasize that the goal is to design and conduct a simulation to find an estimate for the probability that all six people living in an apartment building will get the flu.

Share (consider posting) the steps to designing and conducting a simulation. Student Worksheet 12.2 Simulation Steps lists the steps.

Steps

1. State the problem or statistical/probabilistic question.
2. Define the simple events that form the basis of the simulation.
3. State any underlying conditions that need to be made so the answer to the probabilistic question can be determined.
4. Decide on a model that will be used to match the probabilities. Describe how random numbers will be assigned to match the probabilities described in the problem. Determine what constitutes a trial and what will be recorded.
5. Conduct the first trial.
6. Record the results of the trial.
7. Continue to run trials. Run a large number of trials. Remember to report the result of each trial.
8. Summarize the results of the trials and draw conclusions.

Go through the steps for this simulation using a die or large foam die.

1. State the problem (probabilistic question) so the objective of the simulation is clear. *What is an estimate for the probability all six people living in an apartment building will get the flu?*
2. Define the simple events that form the basis of the simulation. *Infected person randomly visits another person in the apartment building. If a person is randomly visited, they will get the flu, unless they have already had the flu.*
3. State any underlying conditions that need to be made so the answer to the probabilistic question can be determined.

Table 12.1

Trial Number	Who Was Infected (# on Each Roll	Number of People Infected
1	3,4,2,5,3	4
2	6,6,2,6	2
3		

Conditions: Visits are done randomly. Only one person can become infected at a time. Person can infect others for only one day.

4. Decide on a model that will be used to match the probabilities. Describe how random numbers will be assigned to match the probabilities described in the problem. Determine what constitutes a trial and what will be recorded. *Number the people from 1 to 6. Roll a die to simulate the visit by the infected person. (Persons can't visit themselves.) A trial is rolling the die until the flu dies out—a person with the flu visits someone who is immune (already had the flu). The number of people infected will be recorded.*
5. Define and conduct the first trial. *The first roll of the die determines which person was the first person to get the flu. Continue to roll the die until whoever is the current infected person visits an immune person (someone who has already had the flu). That is, roll until a number (other than the infected person's) is repeated. The trial is then over.*
6. Record the results of the trial. *Record the trial number, the results of each roll, and the number of people infected in a table, as shown in Table 12.1.*
7. Continue to run several more trials. Remember to record the result of each trial. *Repeat steps 5 and 6 a large number of times (at least 50 for the class). Give each pair of*

Table 12.2

Number of People Infected	Frequency
2	
3	
4	
5	
6	
Total	

Table 12.3

Number of People Infected	Frequency
2	17
3	33
4	22
5	8
6	2
Total	82

students a die and have them conduct at least five trials and collect the class results in a table.

Explain that an accurate estimate for a probability requires that a large number of trials be conducted (at least 50 for the whole class). Divide the students into groups of two. One person rolls the die and the other records the outcomes in a chart. Ask each group of students to conduct at least five trials.

After the groups have completed at least five trials, collect each group's results in Table 12.2. You are collecting the number of people infected for each trial.

Sample results from a class of 9th graders are shown in Table 12.3.

Option: Demonstrate how to use technology (e.g., ProbSim app on TI-84 Graphing calculator or other rolling die simulator) to collect a large number of trial results.

Analyze the Data

After the simulation has been run for a large number of trials and the results collected in a table, ask the students to answer questions 1 to 4.

1. Fill in Table 12.2 using the class simulation results.
2. Construct a dot plot of the class simulation results.

Possible answer: *Sample results from a 9th grade class in Figure 12.2*

3. What is the most likely number of people living in the apartment building who will get the flu?

Possible answer: *Three people*

4. Add a column to Table 12.3. Label the column Relative Frequency. Complete the relative frequency column in Table 12.4.

Answer: *Based on the example*

Explain that Table 12.4 gives estimates for the relative frequency of various successes (the number of persons who become infected). The relative frequencies for the different number of successes can be thought of as the probability of the number of successes. This table describes a *probability distribution*.

Let X = Number of people infected and P(X) = the probability of x people being infected.

5. What is an estimate for the probability that all six people living in an apartment building will get the flu?

Answer (based on the example in Table 12.5): *0.024, or 2.4%*

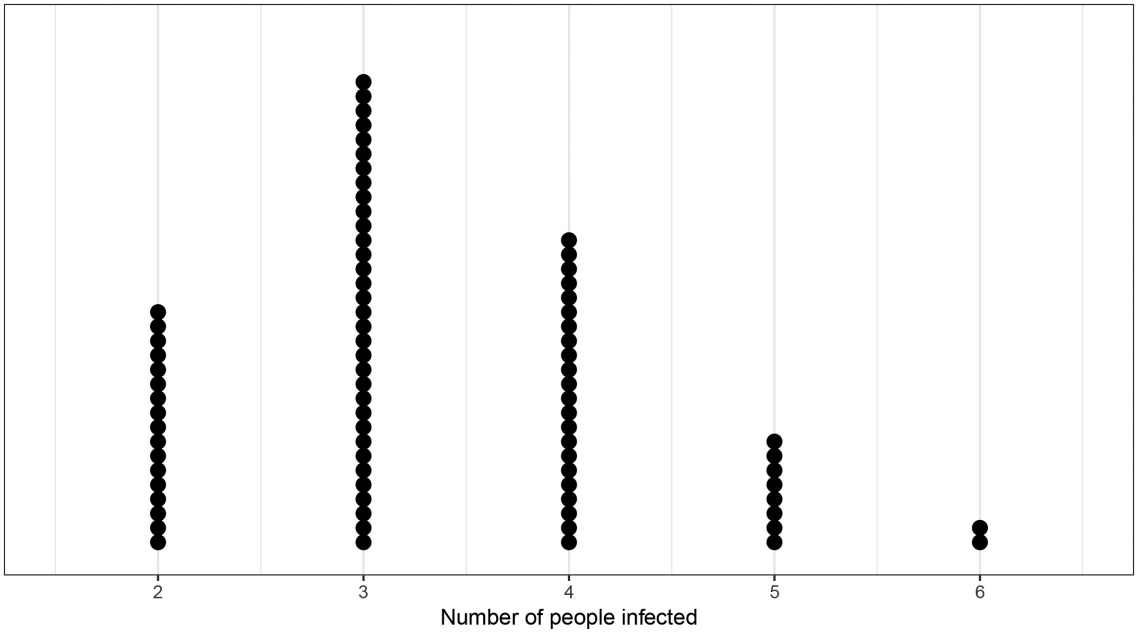

Figure 12.2: Dot plot of the class simulation results

Interpret the Results in the Context of the Original Question

Ask students to answer this question based on the simulation model they designed and conducted.

6. How did you model the spread of the flu in the apartment building? And how did you use this model to find an estimate for the probability that all six people living in the apartment building will get the flu?

Possible answer: *We modeled the spread of the flu by using a six-sided die. Each side of the die represented one person in the apartment building. We rolled the die and recorded the person who got the flu. We continued until a person visited someone with the flu, which caused the flu to die out. We recorded the number of people who got the flu and repeated the simulation a large number of times. After many trials, we were able to estimate the probability of all six people getting the flu as 2.4%*

Summary

To help summarize this simulation, ask your students the following questions:

7. What model could be used if there were eight people in the apartment building?

Table 12.4

Number of People Infected	Frequency	Relative Frequency
2	17	17/82 = 0.207
3	33	33/82 = 0.402
4	22	22/82 = 0.268
5	8	8/82 = 0.098
6	2	2/82 = 0.024
Total	82	1.0

Table 12.5

X	P(X)
2	17/82 = 0.207
3	33/82 = 0.402
4	22/82 = 0.268
5	8/82 = 0.098
6	2/82 = 0.024
Total	1.0

Possible answers: *An eight-sided die, randomly selecting numbers from 1 to 8 from a hat or bag, random number generator on computer or calculator*

8. How do you think the probability of all eight people in an apartment building getting the flu compares with the probability of all six people getting the flu?

Answer: *The probability of eight would be smaller than the probability of six getting the flu.*

Additional Ideas

1. Design and conduct a simulation for the following problem: The chance of contracting strep throat when encountering an infected person is estimated as 0.15. Suppose the four children of a family encounter an infected person. Conduct a simulation to estimate the probability of at least two of the children getting strep throat. State the conditions needed to simulate the problem.
2. A high-school algebra teacher has eight keys, but she never recalls which one fits her office door lock. She tries one key at a time, each time choosing one of the keys at random from her pocket. (All the keys look the same but she does not put a key back in her pocket once she has tried that key.) Conduct a simulation to estimate the probability it will take more than four tries to find the right key.

Exit Ticket

Your math teacher owns 10 ties and randomly chooses a tie to wear to work each school day (not much fashion sense). You notice he sometimes wears the same tie more than once during the week. You wonder if this is likely to happen often, so you decide you would like to find an estimate for the probability he wears the same tie more than once in a five-day workweek. To find this estimate, you design and conduct a simulation.

1. Describe the simple event.

Answer: *Randomly choosing a tie.*

2. Describe a model that would be appropriate to use for the simple event.

Answer: *Number the ties from 1 to 10*

Exit Ticket Cont.

3. Describe a trial and what you would record for each trial.

Answer: *Randomly choose five numbers between 1 and 10. The numbers could be chosen using a 10-sided die, randomly selecting numbers from 1 to 10 from a hat or bag or using the randint(1,10,5) function of the TI-84 graphing calculator or another rolling die simulator.*

Record whether or not a number is repeated.

4. Using the results below, what is an estimate for the probability he wears the same tie more than once in a five-day workweek?

Answer: *An estimate for the probability that he wears the same tie more than once in a five-day workweek is 17/28 = 0.61.*

Table 12.6: Results for the Simulation

Trial Number	Wears Same Tie More Than Once (Y/N)	Trial Number	Wears Same Tie More Than Once (Y/N)
1	Y	15	N
2	N	16	Y
3	N	17	Y
4	N	18	Y
5	N	19	Y
6	Y	20	Y
7	Y	21	Y
8	N	22	N
9	N	23	Y
10	Y	24	N
11	Y	25	Y
12	Y	26	N
13	Y	27	N
14	Y	28	Y

Further Explorations and Extension

Investigating the flu probability problem further.

The simulation gave an estimate for the probability of all six getting the flu. Using formal probability rules, find the exact probability of 2, 3, 4, 5, and 6 people getting the flu. Compare these answers to the simulated probability distribution developed in this lesson.

The probabilities can be calculated in the following manner:

Let X = the number of people infected

P(X=2) = 5/5 * 1/5 = 1/5 = 0.2

P(X=3) = 5/5 * 4/5 * 2/5 = 40/125 = 0.32

P(X=4) = 5/5 * 4/5* 3/5* 3/5 = 180/625 = 0.288

P(X=5) = 5/5 * 4/5* 3/5* 2/5 *4/5 = 480/3125= 0.1536

P(X=6) = 5/5 * 4/5 * 3/5* 2/5 * 1/5 = 120/3125 = 0.0384

Explanation for P(X=3)

5/5 = the probability Person 1 picks another person

4/5 = the probability Person 2 picks another person other than Person 1

2/5 = the probability the third person picks Person 1 or Person 2, which stops the flu.

Answer: *Table 12.7 Probability Distribution*

Table 12.7

X	P(X)
2	0.2
3	0.32
4	0.288
5	0.1536
6	0.0384
Total	1.0

Investigation 13

What Is the Expected Cost to Raise a Child?

Expected Value

Overview

This session begins by introducing the definition of expected value of a random variable. Probability distributions are used to describe and model the behavior of a random variable. The expected value of the probability distribution is calculated and interpreted as the mean of the probability distribution.

This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are formulate a statistical question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level B activity.

This activity is based on lessons from *Probability Models*, published by the American Statistical Association (original copyright by Dale Seymour Publications 1999). *Probability Models* is a module in the ASA Data-Driven Mathematics Project. It is available as a free download *www.amstat.org*.

Instructional Plan

Brief Overview

» Introduce the concept of *expected value*, the notation and connection to the mean of a probability distribution.

» Read and discuss the scenario pertaining to the Nielsen ratings.

» Formulate the statistical question: "If a US family is randomly selected, how much would you expect the cost to be for raising all the children in the family for one year?"

» Find the expected value of the given probability distribution of the number of children under 18 in a family.

» *Optional:* Application of expected value

Introducing Expected Value of a Random Variable

Handout Student Worksheet 13.1 Introducing Expected Value.

During the 2018 Major League Baseball playoffs, one team played in 10 games before losing the division series. In the 10 games, they scored one run in two games, two runs in one game, four runs in four games, and six runs in three games.

1. Complete the frequency table below.

Answer:

Number of Runs	Frequency
1	2
2	1
3	0
4	4
5	0
6	3

2. Construct a dot plot for the runs scored with the variable "runs scored" on the horizontal axis.

Answer: *Figure 13.1*

3. What is the mean number of runs scored per game for the team? Explain how you found the answer and interpret the mean.

Answer: *3.8 runs.*
[1(2) + 2(1) + 4(4) + 6(3)] / 10 = 3.8

Suppose you knew the team scored one run in 20% of the games, two runs in 10% of the games, four runs in 40% of the games, and six runs in 30% of the games, but you didn't know the number of games played.

4. Construct a histogram of the number of runs scored with the vertical axis as relative frequency. How does this graph compare with the dot plot in problem 2?

Possible answer: *The graph in Figure 13.2 is similar, except the vertical axis represents the fraction of games played, rather than number of games played.*

5. Calculate the mean number of runs scored. Explain how you found the answer.

Learning Goal

Understand how to compute and interpret the expected value of a random variable as the mean of the probability distribution.

Mathematical Practices Through a Statistical Lens

MP4. Model with Mathematics

Statistically proficient students can apply mathematics to help answer statistical questions arising in everyday life, society, and the workplace.

Materials

Student worksheets are available at *www.statisticsteacher.org/statistics-teacher-publications/focus.*

» Student Worksheet 13.1 Introducing Expected Value

» Student Worksheet 13.2 Applying Expected Value

» Optional: Student Worksheet 13.3 Application of Expected Value

» Optional: Census Bureau website: *www.census.gov*

» Exit Ticket

Estimated Time

One to two 50-minute class periods

Pre-Knowledge

Students should already be able to find and interpret the weighted mean of a distribution.

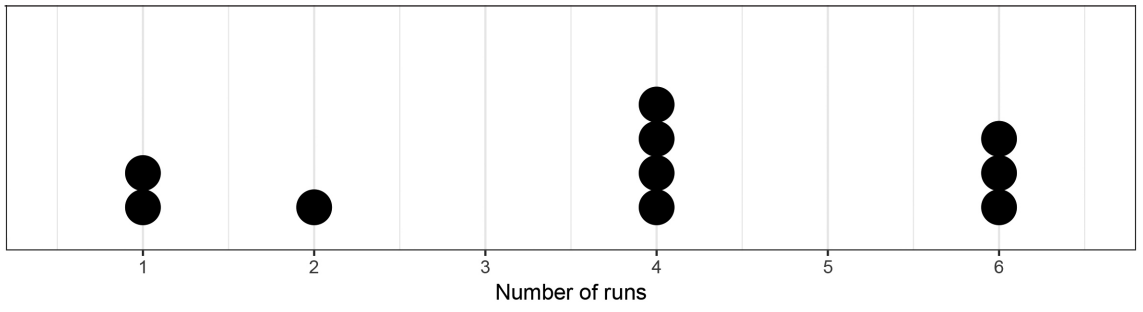

Figure 13.1: Dot plot for the runs scored

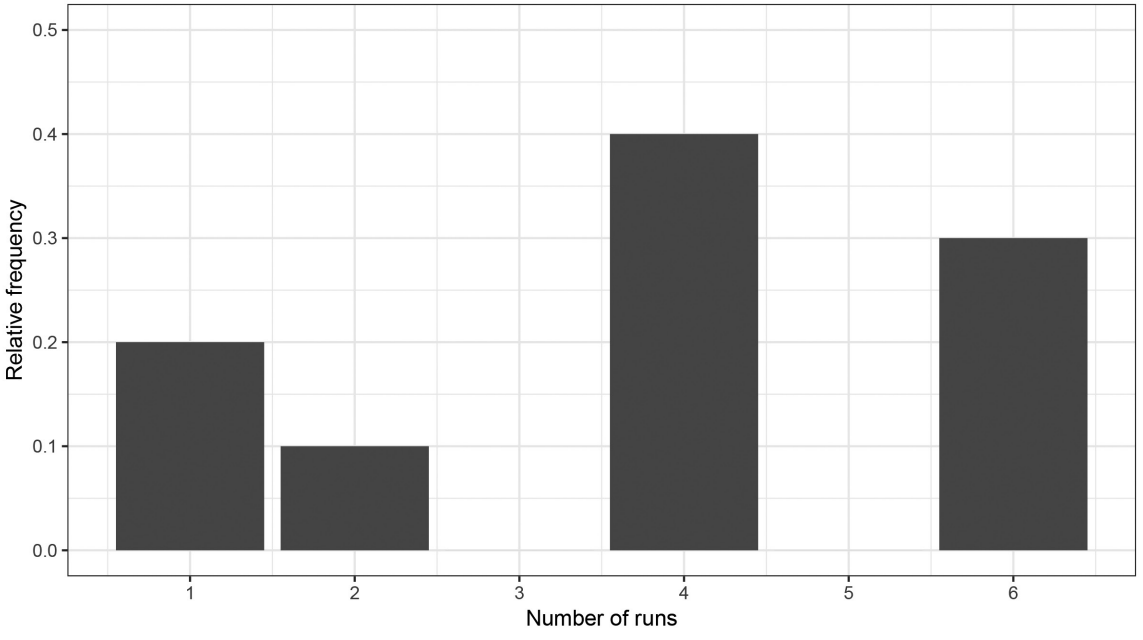

Figure 13.2: Histogram of the number of runs scored

Answer: *3.8 runs.*
1(.20) + 2(.10) + 4(.40) + 6(.30) = 3.8

The team is expected to perform about the same at the start of the next season as they did in the playoffs. That is, the probability of scoring one run in a game is about 20%, scoring two runs in a game about 10%, scoring four runs in a game about 40%, and scoring six runs in a game about 30%.

A mean calculated from a probability distribution—an anticipated distribution of outcomes—is called an expected value.

Let the variable X = the number of runs scored by the team. The variable X is called a random variable.

The probability distribution for a random variable can be displayed in a two-column table, like Table 13.1, for the variable X—the number of runs scored. The symbol P(X) represents the probability of the team scoring X runs in a randomly selected game.

A commonly used symbol for the expected value of X is E(X).

6. Write a symbolic expression for the expected value of X. Recall how you

Table 13.1

X	P(X)
1	0.2
2	0.1
3	0
4	0.4
5	0
6	0.3

Table 13.2

X	P(X)
x_1	p_1
x_2	p_2
x_3	p_3
x_k	p_k

Table 13.3

Number of Motor Vehicles	Relative Frequency (Rounded to 2 Decimal Places)
0	0.09
1	0.34
2	0.37
3	0.14
4	0.06

explained the method used to find the mean of the distribution.

Note: This may be a difficult question, and students may need some scaffolding. One suggestion is to set up a table showing x_1, x_2, x_3, …. x_k under a column labeled X and p_1, p_2, p_3, …. p_k under a column labeled P(x). Then, remind students how they would find the mean.

Possible Answer: *Table 13.2*

Expected value of X could be written $E(X) = x_1p_1 + x_2p_2 + x_3p_3 + \ldots + x_kp_K$

or

$$E(X) = \sum_{i=1}^{k} x_i p_i$$

Application of Expected Value (Optional)

Hand out Student Worksheet 13.2 Applying Expected Value.

Table 13.3 shows the distribution of the number of motor vehicles per US household. A "household" is defined by the US Census Bureau as all persons occupying a housing unit such as a house, an apartment or other group of rooms, or a single room. *Source: www.census.gov*

1. The sum of the relative frequencies is 1.00. Does that mean no US household has more than four cars?

Possible Answer: *A household could have more than four cars, but the relative frequency of such households would round to 0.00.*

Suppose the Department of Energy is planning to select a random sample of households in the US to conduct a survey about reformulated gasoline.

Define a random variable M to be the number of motor vehicles in a randomly selected US household.

2. Find and interpret E(M), the expected value of M.

Answer: *We expect about 1.74 cars per US household. 0(0.09) + 1(0.34) +2(0.37) +3(0.14) + 4(0.06) = 1.74*

3. If the Department of Energy randomly selected 1000 households, how many motor vehicles would we expect these households to have?

Answer: *1.74(1000) = 1740 cars*

Scenario

Have your students read the scenario and answer questions 4 to 11.

Have you and your family ever taken part in a TV or radio rating survey? Maybe you were asked what TV shows you watched or what radio station you listened to on a regular basis. Have you heard of the Nielsen rating?

The Nielsen Corporation is a global marketing research firm. This company was founded in 1923 in Chicago by Arthur C. Nielsen Sr. to give marketers reliable and objective information about the impact of marketing and sales programs. One of Nielsen's best known creations is the Nielsen ratings, a system that measures how many people are watching different TV shows or listening to different radio stations. Nielsen uses statistical sampling to randomly select a representative sample of about 5000 households who agree to be part of the rating estimates. To find out what shows people are watching, meters are installed on all the TV sets in the household. These meters keep track of what TVs are on at any given time and what show the TV set is tuned to. *Source: https://en.wikipedia.org/wiki/Nielsen_Corporation*

4. Why might Nielsen ratings be important information for a TV or radio station?

Possible answer: *Higher ratings mean the station can set higher advertising rates.*

Imagine the television network Nick Jr. is interested in what TV shows people under the age of 18 watch. Network executives could ask the Nielsen Corporation for help determining the most-watched children's TV shows.

The Nielsen Corporation plans on selecting a random sample of families across the US and is particularly interested in families with children under the age of 18. Prior to conducting the survey, researchers at Nielsen find data on the number of children in US families on the US Census Bureau website. According to the US Census Bureau, the number of children under 18 years of age per family in 2010 has a distribution shown in Table 13.4. A "family" is defined as a group of two or more persons related by birth, marriage, or adoption, residing together in a household.

Table 13.4

Number of Children Under 18 in a Family	Percent (Rounded to 1 Decimal Place)
0	55.3
1	19.2
2	16.4
3	8.1
4	1.0

Refer to the distribution of number of children under 18 and answer the following questions.

5. Why do you think the percentage of families with 0 children is so high?

Possible Answer: A large number of families have children older than 18 or no children.

6. The sum of the percentages equals 100%. Does that mean no U.S. families have more than four children?

Possible Answer: *A family could have more than four children, but the relative frequency of such households is tiny and would round to 0.0.*

7. Construct a histogram of the number of children under 18 in US families.

Possible Answer: *Figure 13.3*

8. Find the mean and interpret the mean. Locate the mean on the horizontal scale of the histogram. Is the mean in the center of the graph? Why or why not?

Answer: *The mean is 0.803. The mean is not in the center of the graph. It is much closer to 0 because 0 has such a high frequency of occurrence.*

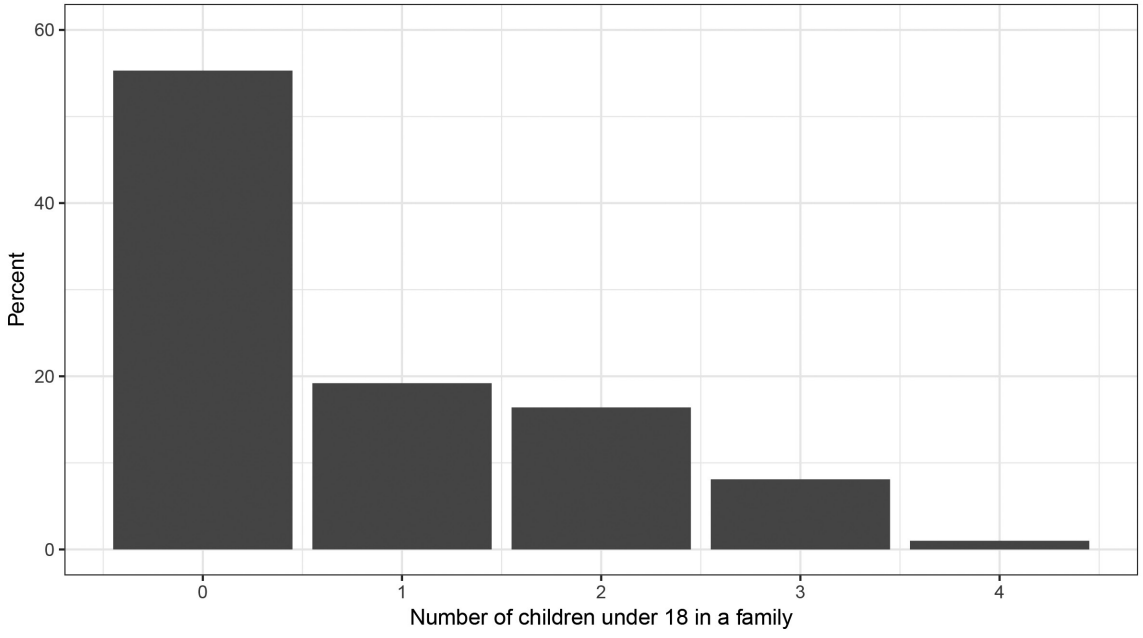

Figure 13.3: Histogram of the number of children under 18 in US families

The A.C. Nielsen Company randomly selects families for use in estimating the ratings of TV shows. Let the variable N represent the number of children under 18 in a randomly selected US family.

9. If A.C. Nielsen randomly selected a family, what is the expected value of N, the number of children under 18 in a randomly selected family?

Answer: *0.803 children per family*

10. How many children in all would we expect to see in a random sample of 2500 families?

Answer: *2500(0.803) is approximately 2008 children.*

11. If Nielsen wants the opinion from at least 1000 children, how many families should be in their random sample?

Answer: *1000/0.803 is approximately 1245 families.*

Formulate a Statistical Question

Suppose families can expect to spend around $13,000 a year to raise a child. Housing, food, child care, clothing, health care, and transportation are some of the expenses.

We want to study a probability distribution for a new random variable C, the cost to raise a child for one year.

Ask your students to consider the statistical question: "If a US family is randomly selected, how much would you expect the cost to be for raising all the children in the family for one year?"

Collect Appropriate Data

12. Using the data from Nielson pertaining to the number of children per family and the cost of raising a child per year, complete the probability distribution (Table 13.5). The first column should contain all the possibilities for C—the cost to raise children in a family for one year.

Answer: Table 13.5

Analyze the Data

13. If a US family is randomly selected, how much would you expect the cost to be for raising all the children in the family?

Answer: 0(0.553) + 13000(0.192) + 26000(0.164) + 39000(0.081) + 52000(0.01) = $10439

Interpret the Results in the Context of the Original Question

14. What was the expected value of N, number of children under 18 in a randomly selected US family?

Answer: 0.803 children per family

15. How is the expected value of C related to the expected value of N?

*Answer: E(C) = 13000*E(N)*

16. Could you have found the expected value of C without building the probability distribution?

Answer: Yes, multiply the cost to raise a child times expected value of N.

Optional: If students need another example, hand out Student Worksheet 13.3 Application of Expected Value.

Scenario

The high-school band is selling raffle tickets to raise money for new uniforms. The winner of the random drawing will receive a necklace designed and made by one of the band parents. The raffle tickets cost $1, and the necklace has a value of $100. The band sells 200 tickets.

Let G represent the amount gained if you buy one ticket.

There are two possible outcomes—you win or lose. If you lose, you have lost your $1, which can be a gain of -1. If you win, your gain would be 100 minus the 1 for the ticket, or a gain of $99.

Table 13.5

C	P(C)
0	0.553
13000	0.192
26000	0.164
39000	0.081
52000	0.01

Table 13.6

Gain/Loss	Probability of Gain/Loss
-$1	199/200
$99	1/200

Table 13.7

Gain/Loss	Probability of Gain/Loss
-$10	190/200
$90	10/200

1. If a person buys one raffle ticket, find the probability distribution for the gain/loss.

Answer: Table 13.6

2. Find the expected gain/loss for a player who buys one raffle ticket.

Answer: Expected value is -0.50, or lose 50¢ for every $1 raffle ticket purchased.

3. What would the expected gain/loss be if a person bought 10 tickets?

Answer: Table 13.7. Expected value is lose $5.

4. What would the expected gain/loss be if a person bought 100 tickets? Can you find the answer without creating a probability distribution?

Answer: Lose $50.

Additional Ideas

Using the Census Bureau website, *www.census.gov*, find the data on the number of TV sets per US household. Using the data, have the students answer the question: "If a US household is randomly selected, what is the expected value of the number of TV sets per US household?"

Exit Ticket

A mobile phone company offers an optional protection that will pay for repairs if the phone is damaged in an accident. The plan costs $50. The retailer has determined the typical cost to repair a broken phone is $150. Let R be the number of repairs a randomly chosen customer will use under this plan. Following is the probability distribution for R.

R	P(R)
0	0.80
1	0.17
2	0.02
3	0.01

1. What is the expected value of R, the number of repairs needed by a randomly selected customer?

Answer: *0.24 repairs*

2. Let C represent the amount it will cost the phone company in repairs for a randomly selected customer. Find the expected value of C.

Answer: *0.24(150), or $36*

3. What is the expected amount of profit the company will make from a randomly selected customer?

Answer: *50−36 = 14. The company is expected to make $14 for each randomly selected customer.*

Further Explorations and Extensions

Have students find and interpret the standard deviation of a probability distribution.

Refer to the probability distribution for the random variable N—the number of children under 18 in a randomly selected US family.

N	P(N)
0	0.553
1	0.192
2	0.164
3	0.081
4	0.01

Possible answer: *The standard deviation of 1.045 is the number of children Nielsen would expect to typically vary per randomly selected family from the expected value of 0.8.*

Investigation 14

How Long Do the Subway Doors Stay Open?

Normal Distribution

Overview

This investigation introduces the Normal distribution as a possible model to describe a sample of times subway doors stay open. The empirical rule is developed and used to help decide if the Normal distribution can be used to model a sample of times. Students will also use the empirical rule to decide what typical door-open times are.

This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level C activity.

Instructional Plan

Brief Overview

» Introduce the empirical rule.

» Develop a statistical question pertaining to the length of time subway doors stay open.

» Decide if the Normal distribution can be used to model the distribution of times subway doors stay open.

Introducing the Normal Distribution

Remind your students that in many of the earlier investigations, we encountered different shapes of distributions. Some were skewed—like the memory test times (Investigation 3)—and others were mound shaped and symmetric—like the length of sample baseball games in 1987 (Investigation 2). The mound-shaped and symmetric distributions are common and key in the study of statistics. Some of these distributions are often referred to as the Normal curve or Normal distribution. Data distributions that are mound shaped and symmetric are often modeled with the Normal curve or Normal distribution.

There are many examples of the application of the Normal distribution. Healthy biological populations such as the heights and weights of animals follow a Normal distribution.

Other examples of Normal distributions include standardized test scores, a person's blood pressure, the weight of packages of cookies, and IQ.

Hand out Student Worksheet 14.1 Normal Distribution.

The example in Figure 14.1 is a distribution of women's heights. The distribution is mound shaped and somewhat symmetric with a mean of 64.6 inches and standard deviation of 2.75 inches. So, we might say the distribution appears approximately Normal.

Another example (Figure 14.2) shows the achievement scores for 200 students at a high school. The distribution is mound shaped and somewhat symmetric with a mean of 75.7 and a standard deviation of 6.1.

Each Normal curve is unique based on the mean and standard deviation, but all have the same shape and properties.

Learning Goals

» Describe a distribution as mound shaped and approaching a Normal distribution.

» Estimate population percentages using the empirical rule.

Mathematical Practices Through a Statistical Lens

MP4. Model with Mathematics

Statistical models build on mathematical models by including descriptions of the variability present in the data.

Materials

Student worksheets are available at *www.statisticsteacher.org/statistics-teacher-publications/focus.*

» Statistical software or app capable of finding summary statistics and constructing a histogram.

» Student Worksheet 14.1 Normal Distribution

» Student Worksheet 14.2 Scenario

» Student Worksheet 14.3 Analyze the Data

» Exit Ticket

Estimated Time

Two 50-minute class periods. One period to introduce the Normal distribution and empirical rule. Another period to determine if a distribution is approximately normal.

Pre-Knowledge

Students should already be able to:

» Use technology to find the mean and standard deviation

» Use technology to construct a dot plot and histogram

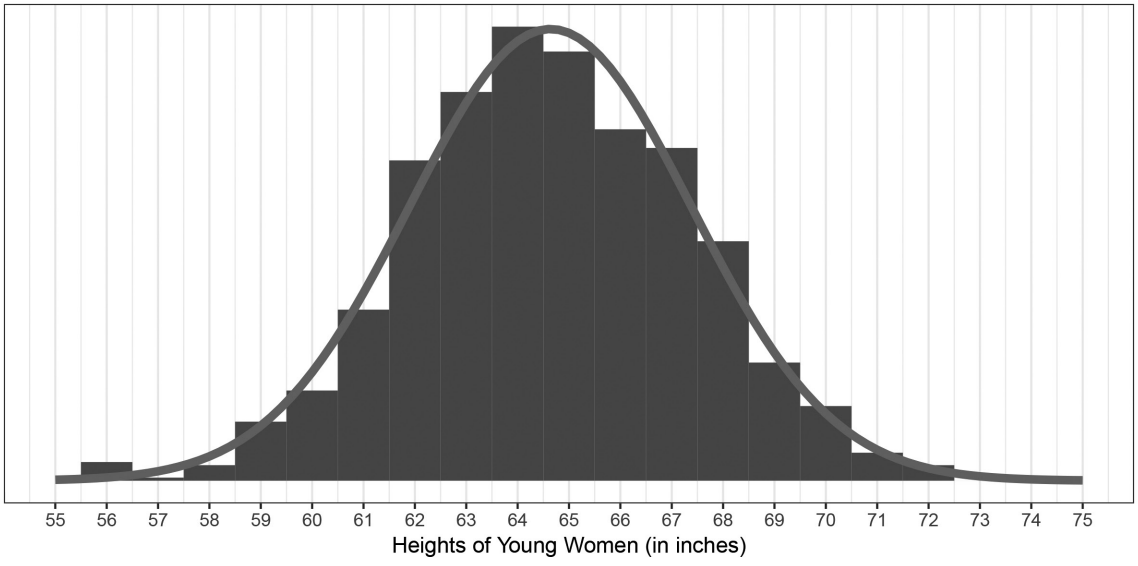

Figure 14.1: Distribution of women's heights

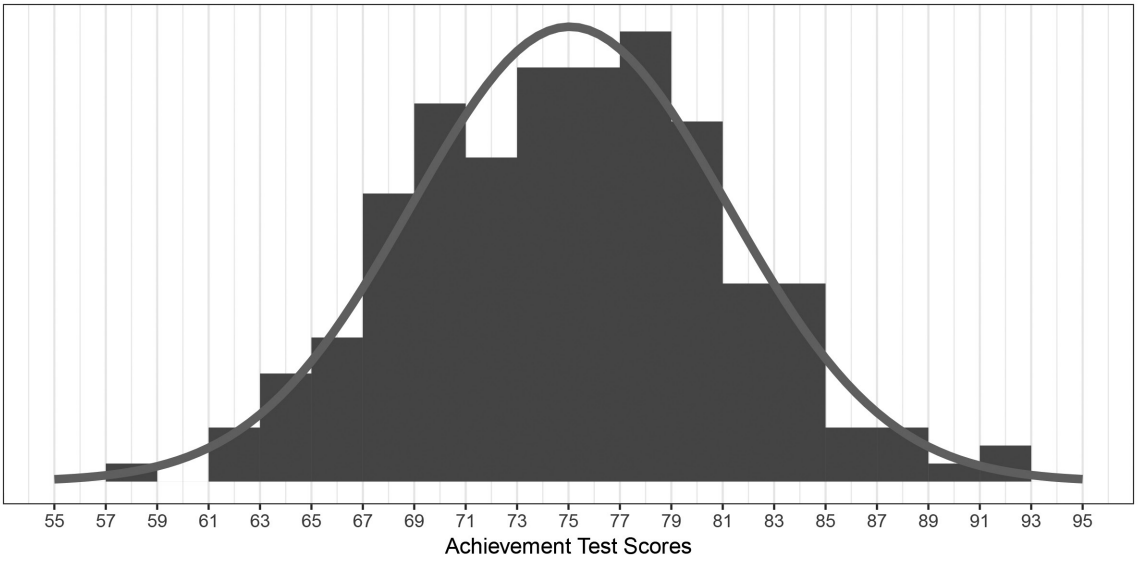

Figure 14.2: Distribution of achievement scores for 200 students at a high school

The proportion of data within one and two standard deviations of the mean is the same for all Normal curves. These proportions form the empirical rule.

The empirical rule states that for a Normal distribution, nearly all the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts, as shown in Figure 14.3:

» Approximately 68% of data falls within one standard deviation of the mean.

» Approximately 95% of data falls within two standard deviations of the mean.

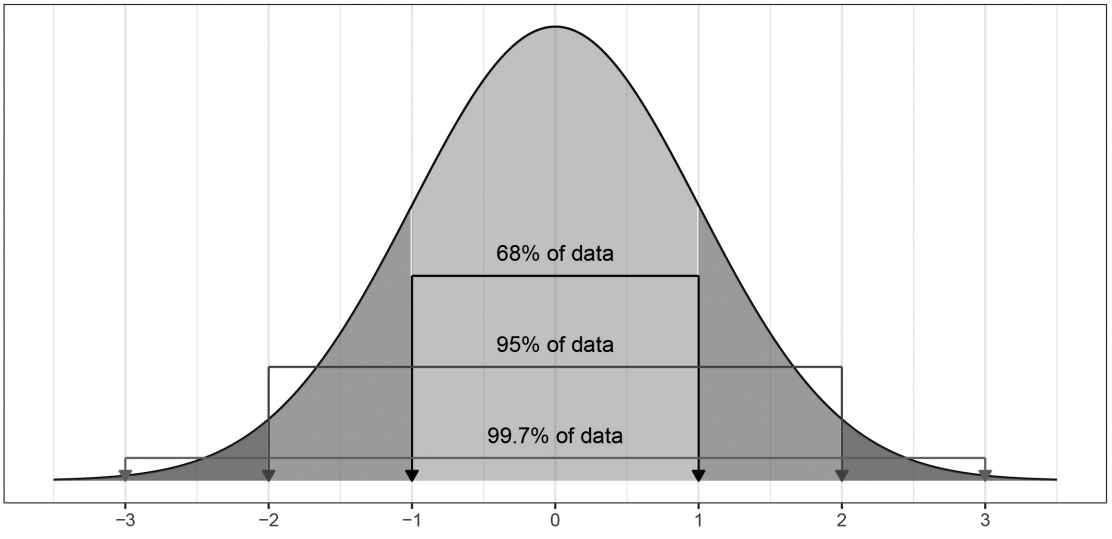

Figure 14.3: The three parts of the empirical rule

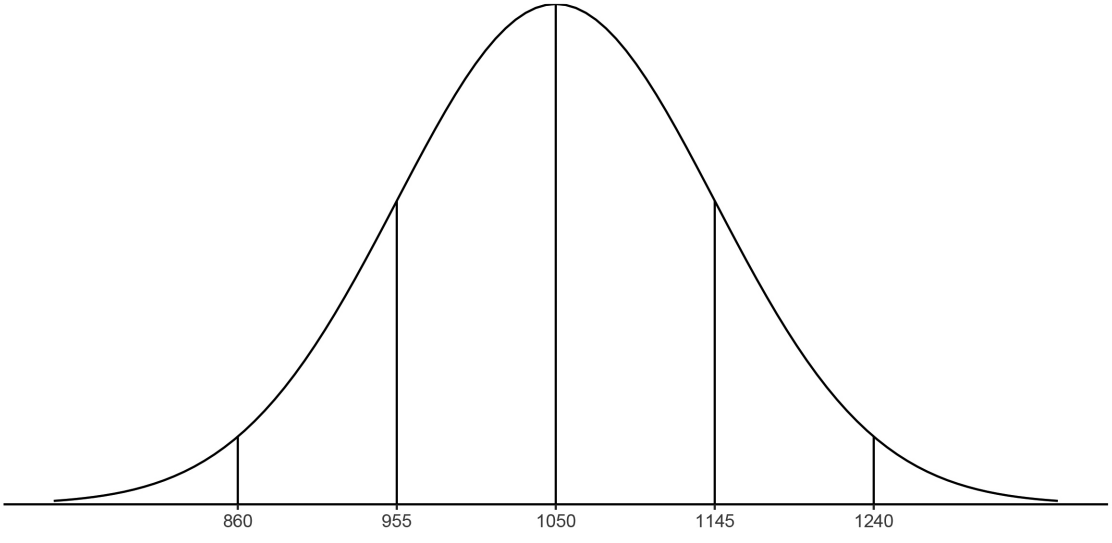

Figure 14.4: Mean and values for one and two standard deviations from the mean

» Approximately 99.7% of data falls within three standard deviations of the mean.

Using the Empirical Rule

Suppose the life of a certain brand of light bulbs can be modeled with the Normal distribution with a mean of 1050 hours and a standard deviation of 95 hours.

1. On the Normal curve in Figure 14.4, add the mean and values for one and two standard deviations from the mean.

Answer: *Figure 14.4*

2. What proportion of bulbs lasts between 955 and 1145 hours?

Answer: *Approximately 68%*

3. What proportion of bulbs lasts between 860 and 1240 hours?

Answer: *Approximately 95%*

4. What proportion of bulbs lasts less than 860 hours?

Answer: *Approximately 2.5%*

5. What proportion of bulbs lasts between 955 and 1050 hours?

Answer: *Approximately 34%*

6. What proportion of bulbs lasts between 1145 and 1240 hours?

Answer: *Approximately 13.5%*

Modeling with the Normal Curve

Hand out Student Worksheet 14.2 Scenario. Give your students time to read the scenario. After they have read the scenario, ask if they have additional questions about the subway door operation.

Scenario

If you live in a large city, it is likely you will encounter a subway or similar mass transportation system. People who live in Washington DC frequently use the Metro. They use the BART in San Francisco, the "L" in Chicago, the MBTA in Boston, the DART in Dallas, and the monorail at Disney World. Mass transportation, however, requires people to make decisions that influence when and where they make their connections. How much time is needed to get to my destination? When should I leave? How long will the doors stay open, or how long will the train wait before leaving? Students in a New York City high school indicated that to get to school on time, they must consider not only when the best time to catch a subway connection is, but also what the chance is a door on the subway will close before they have a chance to get on the train. The students indicated to their teacher that their tardiness to school is often a result of the doors closing before they have a chance to get on board.

In planning the route to school, the following questions were considered by the students:

» When will the subway connection arrive at our location?

» How many people will generally connect at this location?

» What is the expected time the doors will stay open to catch the subway?

Several students indicated that they frequently missed their connection because the doors did not stay open long enough, while other students indicated it rarely happened to them. The students asked a number of questions related to the subway door operation.

» How long do the doors of a subway stay open?

» Are the doors opened and closed automatically or manually?

» Do the doors stay open approximately the same amount of time throughout the day?

» Does the number of people in the subway possibly influence the length of time?

After exploring several factors that might influence the opening and closing of the doors, the students at this school decided to investigate some of these questions through a statistical study. They were particularly interested in the urban myth that the doors on the New York City subway stay open for 30 seconds since they based their decisions of when to catch a subway based on that time.

Formulate a Statistical Question

Point out to your students that the doors on the New York City subway cars are supposed to stay open for 30 seconds according to the urban myth in the scenario. We want to determine if

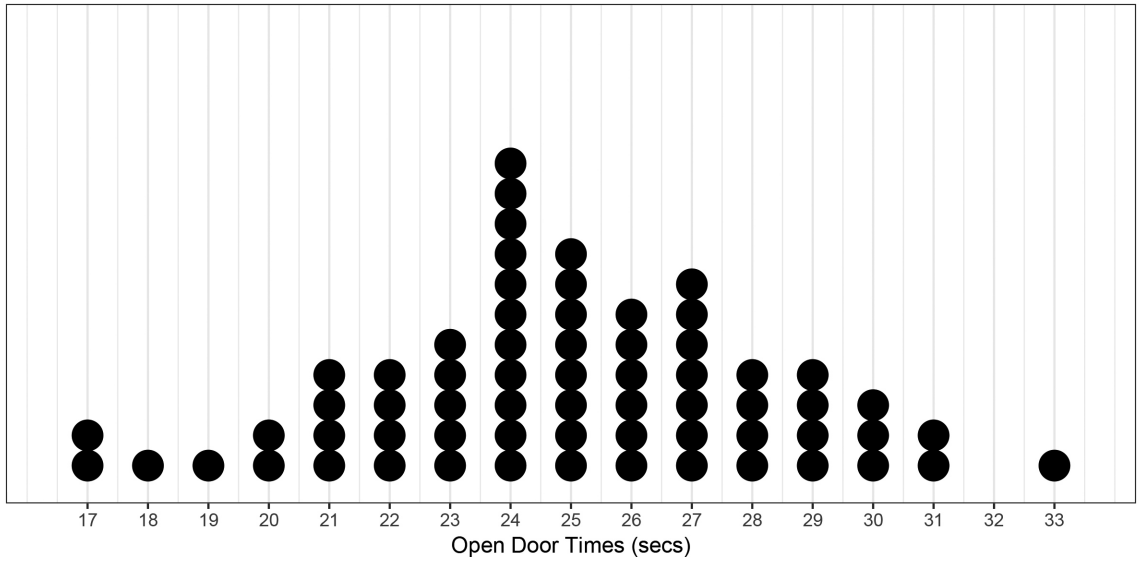

Figure 14.5: Dot plot of the door open times

this myth has validity. Ask your students to consider the statistical questions: "How can we model how long the doors on the F train stay open?" "Is a Normal distribution an appropriate model for the length of time the doors on the F train on the New York City subway route stay open?"

Collection of Data

Ask your students how they would collect times that the doors on the F train stay open. What problems might they encounter in collecting the data?

Answers will vary. They might suggest collecting times at different times of the day, rather than just over the lunch periods, or at different subway stops.

Hand out Student Worksheet 14.3 Analyze the Data.

Discuss with your students how a group of high-school students in Brooklyn, NY, collected data to help answer the statistical questions.

Over a period of 18 school days, a group of students in Brooklyn, New York, recorded how long the doors on the F train stayed open at the subway stop near their school. They collected 65 lengths of time, to the nearest second. Each measurement was taken at approximately the same times each day, usually during three lunch periods.

Below are the data collected:

31 17 19 20 33 29 25 25 26 17 18 29 22 24 24 26 30 27 21 23 28 25 24 21 20 31 28 27 21 24 28 29 30 21 24 27 25 24 23 22 30 26 26 25 25 24 29 24 25 27 27 24 26 22 23 27 22 24 26 28 23 24 25 23 27

Analysis of the Data

Ask your students to answer questions 1 to 4.

1. Construct a dot plot of the door-open times.

Answer: *Figure 14.5*

2. Describe the distribution of the subway door-open times. Include in your description the shape, an estimate for the mean, and an estimate for the standard deviation.

Possible answer: *The distribution is mound shaped with a mean of about 25 sec. and a standard deviation of about 3 sec.*

3. Interpret the mean and standard deviation in this context.

Possible answer: *The mean of 25 seconds is the balance point of the lengths of time deviations on a dot plot. The standard deviation of 3 sec. is the typical length of time a door stays open as measured from the mean.*

4. Using the distribution of subway door-open times, how would you answer the question, "What is the length of time the doors on the F train on the New York City subway route typically stay open?"

Possible answer: *The distribution of subway door-open times centers around 25 seconds. It appears a typical length of time the subway doors are open is approximately 25 seconds. Most of the times were between 21 and 30 seconds. There were only six recorded times when the subway doors stayed open 30 or more seconds.*

After discussing the questions, ask your students to answer questions 5 to 9.

5. To help further understand the distribution of times, complete the frequency chart.

Answer: *Table 14.1: Length of Time Subway Doors Were Open on the F Train*

Table 14.1

Lengths of Time (sec)	Frequency	Lengths of Time (sec)	Frequency
17	**2**	25	**8**
18	**1**	26	**6**
19	**1**	27	**7**
20	**2**	28	**4**
21	**4**	29	**4**
22	**4**	30	**3**
23	**5**	31	**2**
24	**11**	32	**0**
		33	**1**
		Total	**65**

6. What percent of door-open times were:
   a. Less than or equal to 24 seconds?
      Answer: *30/65 = 0.46, or 46%*
   b. More than or equal to 30 seconds?
      Answer: *6/65 = 0.17, or 17%*
   c. Between 22 and 28, inclusive?
      Answer: *45/65 = 0.692, or 69.2%*
7. Convert the frequencies to relative frequencies and record in Table 14.2.
8. Use technology to construct a histogram (bin width of 1 sec.) of the subway door-open times with relative frequency as the

Table 14.2

Lengths of Time (sec.)	Frequency	Relative Frequency	Lengths of Time (sec.)	Frequency	Relative Frequency
17	2	**0.031**	25	8	**0.123**
18	1	**0.015**	26	6	**0.092**
19	1	**0.015**	27	7	**0.108**
20	2	**0.031**	28	4	**0.062**
21	4	**0.062**	29	4	**0.062**
22	4	**0.062**	30	3	**0.046**
23	5	**0.077**	31	2	**0.031**
24	11	**0.169**	32	0	**0**
			33	1	**0.015**
				Total	**1.001**

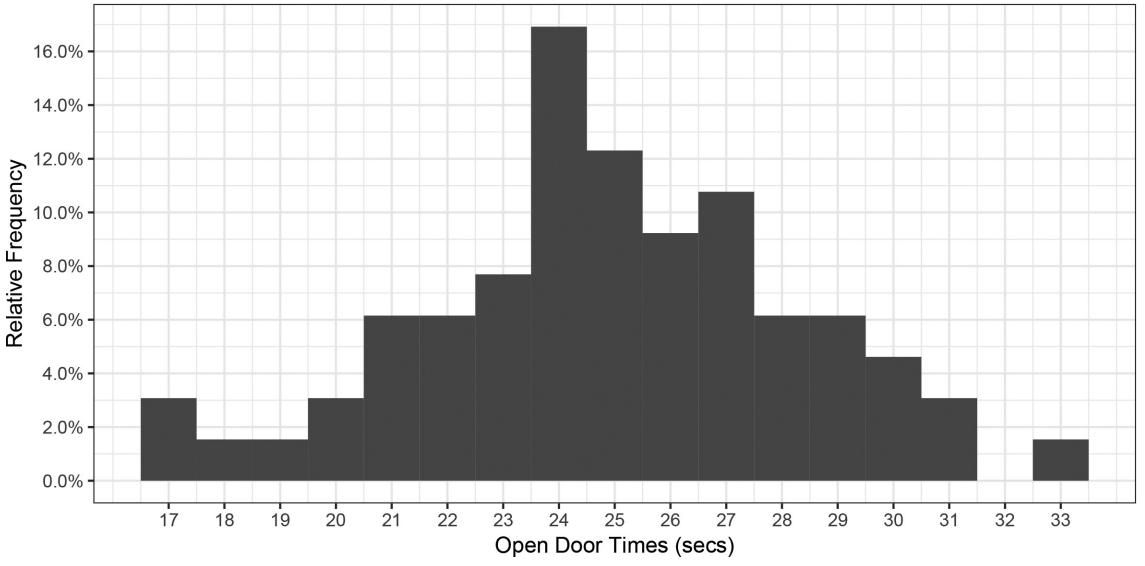

Figure 14.6: A histogram of the subway door open times with relative frequency as the y-axis.

y-axis. Use technology to find the mean and standard deviation.

Note: Students may need help finding the mean and standard deviation with grouped data.

Possible answer: *Figure 14.6. The distribution of subway open door times has a mean and standard deviation of 24.9 sec. and 3.4 sec., respectively.*

9. What are the similarities and differences between the two graphs (dot plot and histogram)?

Possible answer: *Both graphs of the distribution of subway door-open times have the same shape—approximately mound shaped—same center around 25 seconds, and same standard deviation. The only difference is the scale on the y-axis.*

After discussing questions 5 to 9, ask your students to complete Question 10. Have a brief discussion about their responses.

10. Do you think a Normal distribution is an appropriate model for the length of time the doors on the F train on the New York City subway route stay open?

Answers: *Answers will vary, but many students will say the distribution appears to be approximately Normal.*

Explain to your students that we compare some of the properties of a Normal distribution to the distribution of times to further investigate whether the Normal distribution is a good model for the times the subway door stays open.

To do that, let's assume a Normal distribution is an appropriate model with a mean of 24.9 sec. and a standard deviation of 3.4 sec.

As a class, work through questions 11 to 16.

11. On the Normal curve below, locate the mean and the times within one and two standard deviations of the mean by drawing a vertical line through these points.

Answer: *Figure 14.7*

12. Using the empirical rule, what proportion of the data is within one standard deviation of the mean? Show this on the Normal curve.

Answer: *Figure 14.8; approximately 68%*

13. Using the empirical rule, what proportion of the data is within two standard deviations of the mean? Show this on the Normal curve.

Answer: *Figure 14.9; approximately 95%*

14. Using the data table of the relative frequencies, approximately what percent of the open-door times are within one standard deviation of the mean?

Answer: *24.9 − 3.4 = 21.5, 24.9 + 3.4 = 28.3. The sum of the relative frequencies from 22 seconds to about 28 seconds is approximately 0.062+0.077+0.169+0.123+0.092+ 0.108+0.062 = 0.692, or 69.2%.*

15. Using the data table of the relative frequencies, approximately what percent of the open-door times are within two standard deviations of the mean?

Answer: *24.9 − 2*3.4 = 18.1, 24.9 + 2*3.4 = 31.7. The sum of the relative frequencies from 18 to 32 is approximately 0.693+.015+.015+.031 +.062+.062+.046+.031= 0.955, or 95.5%.*

16. How do the proportions you found in questions 14 and 15 compare with the empirical rule percentages based on the Normal curve as a model for the times the subway doors are open?

Possible answer: *The percentages are very close to the theoretical percentages.*

Interpret the Results in the Context of the Original Question

Give students time to answer questions 17 to 19.

17. Do you think a Normal distribution is an appropriate model for the length of time the doors on the F train on the New York City subway route stay open?

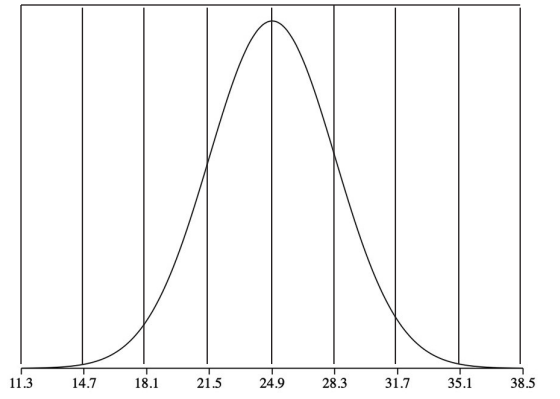

Figure 14.7: The mean and times within one and two standard deviations of the mean

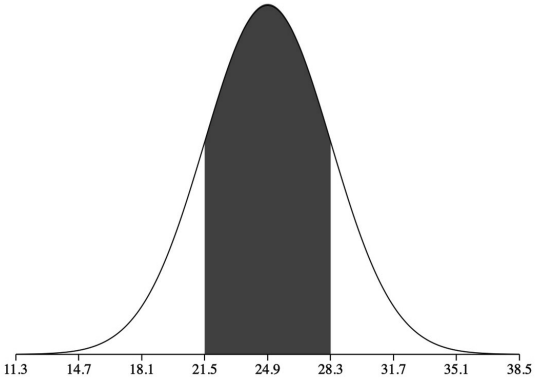

Figure 14.8: The proportion of data within one standard deviation of the mean

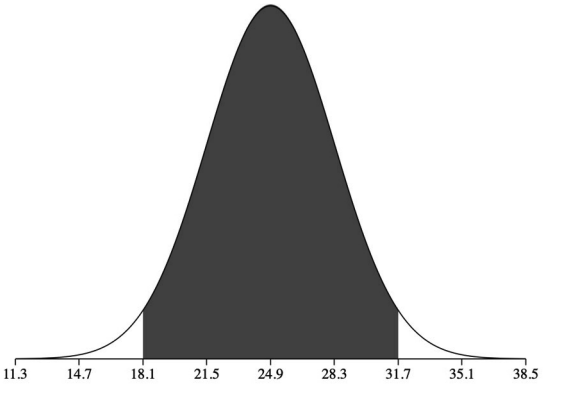

Figure 14.9: The proportion of data within two standard deviations of the mean

Answer: *The distribution is mound shaped and symmetric, and the percent of data within one and two standard deviations of the mean approximately equal the percentages found in the empirical rule.*

18. Use the model (Normal curve) you have created and decide if the urban myth that the subway doors typically stay open for 30 seconds is true.

Possible answer: *It seems unlikely the doors stay open for 30 seconds. Out of the 65 times collected, the doors stayed open 30 or more seconds only six times, or less than 10% of the time. It appears the length of time the doors stay open is approximately a Normal distribution with a mean of about 25 seconds and a standard deviation of about 3.5 seconds. Thirty seconds is well outside what we would expect.*

19. Another method to determine whether there are outliers is to use the standard deviation. If the distribution is mound shaped (approximately Normal), then any data point more that two standard deviations from the mean can be considered an outlier. Using this definition, are there any outliers in the subway door-open times? Is the urban myth of the doors staying open 30 seconds an outlier?

Possible answer: *Any value less than 18.1 seconds or three values at 18, 17, and 17. Any value greater than 31.7 or one value at 33. 30 seconds is not an outlier, but it is still outside the typical length of time the doors stay open.*

Additional Ideas

In lieu of using the data presented in this lesson, students could do the following:

» Collect data from the internet/newspaper on the price of a particular model of new car and explore the Normal distribution as a model for the distribution.

» Collect data from the internet/newspaper on the price of a particular model of used car and explore the Normal distribution as a model for the distribution.

» Collect the times to go through the lunch line in the school cafeteria and explore the Normal distribution as a model for the distribution.

» Collect data from the internet/newspaper on the batting averages for major league baseball players and explore the Normal distribution as a model for the distribution.

» Collect data from the internet on the length of Old Faithful's eruption times and explore the Normal distribution as a model for the distribution.

Exit Ticket

Alicia, a senior in high school, would like to find out how her SAT score compares with other seniors who took the SAT. She decided to investigate the question: "What is a typical score on the math portion of the SAT test?"

The College Board reported in 2016 that approximately 1.7 million high-school students took the SAT. The scores on the math portion of the test were approximately Normally distributed with a mean score of 513 and standard deviation of approximately 118.

1. Draw a Normal curve and mark the mean and one standard deviation and two standard deviations from the mean.

Answer: *Figure 14.10*

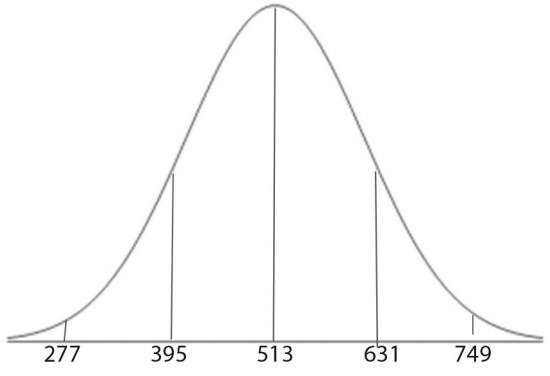

Figure 14.10: Mean with one and two standard deviations from the mean

2. What proportion of scores is within one standard deviation of the mean?

Answer: *Approximately 68%*

3. What proportion of SAT scores is between 277 and 749?

Answer: *Approximately 95%*

4. If Alicia had a score of 631, approximately what percent of the students did she do better than?

Answer: *Alicia scored higher than approximately 84% of the students who took the SAT test.*

Further Explorations and Extensions

The term "Normal curve" came from the study of the errors made in astronomical observations and other scientific observations. Abraham de Moivre, an 18th-century statistician and gambling consultant, was often asked to make lengthy computations, like the probability of observing at least 60 heads in 100 coin tosses. Because this calculation would be difficult, de Moivre observed as the number of tosses increased, the distribution of the number of heads became more and more mound shaped, symmetrical, and smoothly curved. If he could find a function for this curve, he could find the probability of getting 60 or more heads out of 100 coin flips with much less difficulty.

What we now call the Normal curve is what de Moivre discovered around 1733. The curve may also have been discovered separately by the astronomer and mathematician Pierre Laplace in 1786. A different derivation of the formula was presented in 1809 by Karl Friedrich Gauss, hence the Normal curve is often called the Gaussian curve.

Summarized from *https://onlinestatbook.com/2/normal_distribution/history_normal.html* and *https://sydney.edu.au/stuserv/documents/maths_learning_centre/normal2010web.pdf*

The Normal curve is mound shaped and symmetrical. One form of the equation is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ is the population mean

σ is the population standard deviation

π is the constant pi

e is the constant, Euler's number, 2.718281828 …