Students learn that in many situations they require data from samples to estimate characteristics (parameters) of large populations. The reason may be that the population is too large to permit the student to gather data from every subject. For example a national opinion poll about a political issue. Sometimes the act of collecting the data ruins the item being studied, as when crash-testing automobiles. Understandably, to get reliable results, one must minimize the sources of bias. Students discuss the potential for bias in the wording of survey questions. Later in the section they do an activity which shows the need for the random selection of subjects to reduce bias in statistical studies.

For additional information, see the Math Notes box in Lesson 9.1.1. Additionally, problem 9-3 explains types of bias found within survey questions.

Examples

Determine whether each situation describes a sample or census. If the situation describes a sample, discuss the sampling technique and potential sources of bias.

a. The Chief of Police calls in his five newest officers to get their opinion on new requirements for department-wide promotions.
   
   Answer: The Chief is using a sample, which is not likely representative of the population affected by the new requirements. The newest officers would likely want to please the Chief and would echo his support or dislike of the new requirements.

b. A manager compares the annual sales totals for 12 store locations to rank their performance.
   
   Answer: This is a census. The manager has information from all 12 of the stores he is interested in comparing.

c. Shoppers are invited to fill out a questionnaire about their shopping experience at the cash register. The store manager uses the results in a report to corporate headquarters to demonstrate the level of customer satisfaction at her store.
   
   Answer: The manager is using a voluntary response sample (or “convenience sample”), which would not represent all customers very well. A sample like this tends to over-represent those with the strongest opinions.

d. The student council is trying to determine how much space they will need for the prom. Carmine, the Junior Class President, walks around during lunchtime at her school with a clipboard, asking, “Are you planning on going to the best prom ever?”
   
   Answer: Carmine is not taking a random sample. She is more likely to talk to students she knows, and they, as a group, are likely to be biased about the prom. There may also be a strong desire to please the interviewer in this situation. The wording of the question itself is biased, using “best prom ever” to describe the dance.
Problems

Determine whether each situation describes a sample or census. If it is a sample, discuss the sampling technique and potential sources of bias.

1. A math teacher wants to determine whether playing classical music during testing benefits high school math students. He plays classical music to half of his classes while they test, and compares their scores to the other half of his classes, who tested without the music.

2. A sports news reporter wants to know the win-loss record of the local high school girls lacrosse team, so he looks at the league website and sees that they have 8 wins and 5 losses.

3. A battery manufacturer wants to monitor the durability and peak voltage of its products. A machine is programmed to select every 1000th battery from the production line and subject it to a series of tests.

4. Scott is doing a report on the risk factors for cancer. He asks all of the members of his PE class: “Do you have a family history of cancer?” and “Do you eat 5 servings of fruits and vegetables each day?”

Answers

1. Because the math teacher is using his data to make a statement about all high school math students (the population), his classes represent a sample of those students. His students would have many things in common, like where they live and how they were taught, which other high school math students do not have. They would be a poor representation of the population. The teacher may also be biased in grading the exams to obtain the results he favors.

2. This is a census. The reporter has access to every game outcome for the team he is researching and is not using the information to draw conclusions about other teams or the sport of lacrosse in general.

3. The batteries tested represent a sample of those being made. They might be a reasonable representation of the population of batteries. However, if the 1000th battery is usually made at the same time every day, or corresponds to some other variable like periodic machine maintenance, it might be subject to bias.

4. Scott’s PE class represents a sample of people or students. However this is a convenience sample and cannot be considered a reliable representation of the population of people or students. The question order may cause bias in the results as people being reminded of a disease like cancer would likely want to appear concerned enough to be eating a healthy diet with lots of fruits and vegetables. Members of a PE class may be more inclined to be thinking about health issues such as diet.
Two ways to gather information for statistical studies are observational studies and experiments. Observational studies seek to collect data about subjects without changing them, while experiments impose treatments (variables of interest) upon the subjects. Observational studies are generally easier to perform but cannot demonstrate cause-and-effect relationships because it is not possible to separate the many effects of all the lurking variables (those outside of what is being observed directly).

For additional information, see the Math Notes box in Lesson 9.2.1.

Examples

For each question below:

- Does a census or a sample make more sense in this situation?
- Should an observational study or an experiment be carried out to answer the question?
- If an observational study is suitable and asks for an association between variables, discuss the effects of possible lurking variables.
- If an experiment would be suitable, outline an experimental design.

a. Is there a relationship between drinking green tea and longevity?
   Answer: Because it would be impossible to find every green tea consumer this study would need to rely on sampling. Green tea would not be a difficult treatment to administer in an experiment but because life-spans are being measured, an experiment would be impractical. An observational study may find a relationship between these two variables but would be unable to establish cause and effect because of the many possible lurking variables. Perhaps people who choose to drink green tea also choose to be more health conscious in other ways such as smoke less than the general population. They may also tend to be of higher income and have better access to health care. Hence their longevity causes could really be smoking less and having better health care, not from drinking green tea.

d. What is the mean verbal SAT score in the state of Colorado for a given year?
   Answer: This information is easily found by a census.

c. What percentage of razorback clams show signs of stress from acidic ocean water?
   Answer: One could not find every clam for study and the stress testing of the clams may harm them, so a sample would be appropriate. It would be important to gather razorback clams from a variety of locations as environmental conditions may vary considerably.

d. Would brighter classroom lighting increase exam scores?
   Answer: This would require using a sample of students in a controlled experiment. One could take student volunteers and give them the same instruction on a topic, then randomly assign them to classrooms with different lighting levels for the same exam. The mean test scores from each group could be compared.
Problems

For each question:

• Does a census or a sample make more sense in this situation?
• Should an observational study or an experiment be carried out to answer the question?
• If an observational study is suitable and asks for an association between variables, discuss the effects of possible lurking variables.
• If an experiment would be suitable, outline an experimental design.

1. Is there a relationship between the temperature of water and its electrical resistance?
2. What is the mean distance students commute to school in your county?
3. What percentage of car crashes causing injury occurred when at least one driver was impaired by drugs or alcohol?
4. Is there an association between baldness and being colorblind?

Answers

1. One would have to work with samples of water in this situation. This question would best be answered as an experiment. Begin with water from the same source so its other variables like impurities remain constant and divide it into several like containers. Randomly assign different temperatures to the water containers and measure the electrical resistance between two wires inserted an equal distance apart in each container, then compare results.

2. For most counties, surveying every student about commute distances is not possible, so a random sample of students could be used to represent the population. This is clearly an observational study because the question seeks to determine how students commute without imposing any changes upon them.

3. This kind of information is routinely collected by law enforcement, so census data is likely available.

4. This study would require samples from the population. An observational study is the only reasonable alternative because an experiment would require imposing baldness or colorblindness upon subjects. A lurking variable here is gender. Men are more often colorblind and more often bald than women, so if the study included both genders then it would likely find an association.
When the frequencies in a frequency table are divided by the total number of data points, the frequencies become ratios, and the frequency table becomes a relative frequency table. Histograms show the information from frequency tables and relative frequency histograms display information from relative frequency tables.

In this section histograms or their associated tables are used to translate ranges of data into percentages.

Note: Many of these lessons require the use of a graphing calculator. Instructions for using a TI-83/84+ calculator can be found at cpm.org or studenthelp.cpm.org.

Example 1

Forty applicants interviewed for an administrative assistant position. Each applicant was given a typing test and earned a score in words per minute (wpm). Their results are shown in the table.

<table>
<thead>
<tr>
<th>Typing score (wpm)</th>
<th>Frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 to 39</td>
<td>1</td>
<td>( \frac{1}{40} = 0.025 )</td>
</tr>
<tr>
<td>40 to 49</td>
<td>6</td>
<td>( \frac{6}{40} = 0.15 )</td>
</tr>
<tr>
<td>50 to 59</td>
<td>11</td>
<td>( \frac{11}{40} = 0.275 )</td>
</tr>
<tr>
<td>60 to 69</td>
<td>16</td>
<td>( \frac{16}{40} = 0.40 )</td>
</tr>
<tr>
<td>70 to 79</td>
<td>5</td>
<td>( \frac{5}{40} = 0.125 )</td>
</tr>
<tr>
<td>80 to 89</td>
<td>1</td>
<td>( \frac{1}{40} = 0.025 )</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>1.00</td>
</tr>
</tbody>
</table>

a. Create a frequency table, histogram, relative frequency table, and relative frequency histogram.

b. What percentage of applicants typed 40 to 49 wpm?

c. What percentage of applicants typed at least 60 wpm?

d. If 50 wpm is the minimum job qualification, what percentage of applicants did not qualify?

Answers:

a. 15% (from the relative frequency table or histogram)

b. 0.40 + 0.125 + 0.025 = 0.55 = 55%

c. 0.025 + 0.15 = 0.175 = 17.5%
Example 2

A random sample of 35 cars was taken from a car rental agency, and the highway mileage of each car was measured in miles per gallon (mpg).

<table>
<thead>
<tr>
<th>Highway Mileage (mpg)</th>
<th>Frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 up to 27</td>
<td>2</td>
<td>$\frac{2}{35} \approx 0.0571$</td>
</tr>
<tr>
<td>27 up to 30</td>
<td>3</td>
<td>$\frac{3}{35} \approx 0.0857$</td>
</tr>
<tr>
<td>30 up to 33</td>
<td>1</td>
<td>$\frac{1}{35} \approx 0.0286$</td>
</tr>
<tr>
<td>33 up to 36</td>
<td>5</td>
<td>$\frac{5}{35} \approx 0.1429$</td>
</tr>
<tr>
<td>36 up to 39</td>
<td>13</td>
<td>$\frac{13}{35} \approx 0.3714$</td>
</tr>
<tr>
<td>39 up to 42</td>
<td>9</td>
<td>$\frac{9}{35} \approx 0.2571$</td>
</tr>
<tr>
<td>39 up to 45</td>
<td>2</td>
<td>$\frac{2}{35} \approx 0.0571$</td>
</tr>
<tr>
<td>total</td>
<td>35</td>
<td>1.00</td>
</tr>
</tbody>
</table>

c. $\approx 0.0571 + 0.0857 = 0.1428 = 14.28\%$

d. Counting the cars from either end of the mileage distribution, you would find that the 18th car would be in the middle. The 18th car would also be in the 36 up to 39 mpg group.

e. $\approx 0.2571 + 0.0571 = 0.3142 = 31.42\%$
THE NORMAL PROBABILITY DENSITY FUNCTION AND PERCENTILES

There are many situations when a relative frequency histogram will closely resemble a bell-shaped curve. Sometimes an equation called the normal probability density function is fitted to the bell-shaped relative frequency histogram. This function is commonly called the “normal distribution” and can be used to find percentages and probabilities associated with what is being measured. The normal probability density function is also used to find percentiles. Percentiles identify the measurement below which a specified percentage of data points are found. For example, if the 90th percentile for men’s heights were 72 inches, 90% of men would be shorter than 72 inches.

For additional information, see the Math Notes box in Lesson 9.3.3.

Example 1

Half liter soda bottles are filled with cola at a bottling plant. The equipment used to fill the bottles is not capable of putting exactly 500 ml in each bottle, so it is adjusted such that the amount of cola placed in the bottles follows a normal distribution with a mean of 502 ml and a standard deviation of 1.8 ml.

a. What proportion of bottles will have more than 502 ml of cola in them?

b. What proportion of bottles will have between 500 and 502 ml in them?

c. What proportion of bottles are under-filled (hold less than 500 ml)?

d. What percentile corresponds to 503 ml?

e. What percentile corresponds to 505 ml?

Answers:

a. normalcdf(502, 10^99, 502, 1.8) = 0.500
   One may recognize that 502 ml is the mean and half (0.500) of the probability in a normal distribution is always above and below the mean.

b. normalcdf(500, 502, 502, 1.8) = 0.367

c. normalcdf(−10^99, 500, 502, 1.8) = 0.133

d. normalcdf(−10^99, 503, 502, 1.8) = 0.711 or the 71st percentile

e. normalcdf(−10^99, 505, 502, 1.8) = 0.952 or the 95th percentile
Example 2

Commercial cherry trees produce fruit according to a normal distribution with an average of 90 kg of cherries per tree and a standard deviation of 11 kg per tree.

a. What is the probability that a tree produces at least 95 kg of cherries?
b. What is the probability that a tree yields between 85 and 105 kg of cherries?
c. What percentile would a harvest of 80 kg be for a single tree?

Answers:

a. \( \text{normalcdf}(95, 10^{99}, 90, 11) = 0.325 \)
b. \( \text{normalcdf}(85, 105, 90, 11) = 0.589 \)
c. \( \text{normalcdf}(-10^{99}, 80, 90, 11) = 0.182 \) or the 18\(^{th} \) percentile

Problems

1. Concrete is loaded into trucks for a large construction project according to a normal distribution. The mean amount loaded is 230 cubic feet with a standard deviation of 7 cubic feet. Concrete trucks are rated by the number of cubic yards they can carry, and 9 cubic yards is a common maximum load. Hence the expression “the whole nine yards.”

a. What proportion of the trucks will be loaded with between 220 and 240 cubic feet?
b. What proportion of trucks are loaded with more concrete than the 9 cubic yard maximum. (3 feet = 1 yard)
c. 235 cubic feet is what percentile load?
d. If a particular job requires at least 220 cubic yards of concrete, what is the probability there will not be enough concrete in a single truck?

2. Assume the annual snowfall in the village of Thalhammer follows a normal distribution with a mean of 185 cm and standard deviation of 36 cm.

a. Less than 100 cm of snow in a year is considered drought conditions in Thalhammer. What is the chance that a particular year will have drought conditions?
b. The Rivera Valley lies below Thalhammer. Residents of Rivera Valley are placed on a flood-warning from snowmelt in the spring if annual snowfall in Thalhammer is greater than 250 cm. What is the probability the Rivera Valley residents will be placed on a flood-warning this spring?
c. What percentile corresponds to 160 cm of annual snowfall in Thalhammer?
Answers

1.  a.  \( \text{normalcdf}(220, 240, 230, 7) = 0.847 \)
   
   b.  \( (3\text{ ft})(3\text{ ft})(3\text{ ft}) = 27\text{ ft}^3 \), so 27 cubic feet per cubic yard.  \( (9)(27) = 243 \text{ cubic feet} \)
      \( \text{normalcdf}(243, 10^{99}, 230, 7) = 0.032 \)
   
   c.  \( \text{normalcdf}(-10^{99}, 235, 230, 7) = 0.762 \) or the 76\(^{th}\) percentile
   
   d.  \( \text{normalcdf}(-10^{99}, 220, 230, 7) = 0.077 \)

2.  a.  \( \text{normalcdf}(-10^{99}, 100, 185, 36) = 0.0091 \)
   
   b.  \( \text{normalcdf}(250, 10^{99}, 185, 36) = 0.035 \)
   
   c.  \( \text{normalcdf}(-10^{99}, 160, 185, 36) = 0.244 \) or the 24\(^{th}\) percentile
1. In the rectangle $ABCD$ at right, the area of the shaded region is given by $\frac{\pi lw}{6}$. If the area of the shaded region is $7\pi$, what is the total area, to the nearest whole number, of the unshaded regions of the rectangle $ABCD$?
   a. 14  b. 15  c. 20  d. 22  e. 25

2. Consider the following equations:
   
   
   
   
   
   
   
   
   
   
   
   If $p$ is a negative integer, what is the ordering of $a$, $b$, and $c$ from least to greatest?
   a. $c < a < b$  b. $a < c < b$  c. $b < a < c$  d. $a < b < c$  e. $c < b < a$

3. The figure at right represents six offices that will be assigned randomly to six different employees, one employee per office. If Maryanne and Ginger are two of the six employees, what is the probability that they will be assigned an office indicated with an *?
   a. $\frac{1}{6}$  b. $\frac{1}{8}$  c. $\frac{1}{15}$  d. $\frac{2}{15}$  e. $\frac{1}{30}$

4. Raul needed wire pieces 7 inches long. He cut as many as he possibly could from a wire 6 feet long. What is the total length of the wire that is left over?
   a. 2 inches  b. 3 inches  c. 4 inches  d. 5 inches  e. 8 inches

5. The $n^{\text{th}}$ term of a sequence is defined to be $5n + 2$. The $35^{\text{th}}$ term is how much greater than the $30^{\text{th}}$ term?
   a. 5  b. 18  c. 25  d. 36  e. 40

6. Matilda remembers only the first four digits of a seven-digit phone number. She is certain that none of the last three digits is zero. If she dials the first four digits, then dials the last three digits randomly from the non-zero digits, what is the probability that she will dial the correct number?

7. Let $a \bullet b$ be defined as $\frac{1}{a} + b$ where $a \neq 0$. What is the value of $6 \bullet 7$?
8. If $4xy + 1 = 1$, what is the value of $xy$?

9. Eight consecutive integers are arranged in order from smallest to the largest. If the sum of the first four integers is 206, what is the sum of the last four integers?

10. If the points $A(4, 1)$, $B(4, 8)$, and $C(-3, 8)$ form the vertices of a triangle, what is the area of the triangle?

**Answers**

1. C
2. D
3. C
4. A
5. C
6. $\frac{1}{729}$
7. $\frac{43}{6}$
8. 0
9. 222
10. 24.5