Session 2

Data Organization and Representation

Key Terms for This Session

Previously Introduced

- data
- quantitative data
- variation

New in This Session

- cumulative frequency
- cumulative frequency table
- discrete data
- distribution
- frequency
- frequency bar graph
- frequency table
- interval
- line plot
- median
- mode
- relative frequency
- relative frequency bar graph

Introduction

In the previous session, you explored measurement and variation. You learned that there is almost always variation in statistical data, and you looked at potential sources of the variation, including random error and bias. You may also recall that there are four components to statistical problem solving:

1. Ask a question.
2. Collect appropriate data.
3. Analyze the data.
4. Interpret the results.

In this session, we will focus on the last two steps of this process—analyzing data and interpreting results. We will explore different graphical and tabular representations and their usefulness in helping us analyze the data. We'll also begin to look at how particular values can summarize or "best represent" the entire data set, a topic that will be discussed in more detail in later sessions. **[See Note 1]**

Learning Objectives

In this session, you will learn how to do the following:

- Organize data in a line plot and a frequency table
- Organize data in a cumulative frequency table
- Use intervals to answer a statistical question
- Determine the median of a set of data
- Determine relative frequencies and create bar graphs of your data

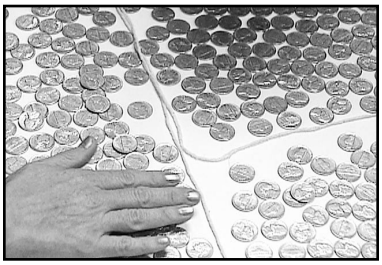
---

Note 1. Materials Needed: large poster board, collection of nickels (about 100), magnifying glasses (if needed to see the mint marks on the coins), string, sets of 17 half-ounce boxes of raisins (one set per group or for each individual working alone)—all 17 should be from the same brand, but each set should be different; e.g., 17 boxes of Brand X, 17 boxes of Brand Y, etc.

Part A: Patterns in Variation (10 min.)

To answer a statistical question, we collect data, which consist of measurements of a variable. When the measurements we collect differ from each other, variation exists. We can examine this variation in several ways to find interesting answers to the statistical questions we pose.

For example, we can look at the distribution of a data set—the different values and how often each occurs—to draw out informative patterns in the variation. Tabular and graphical representations of the data can help to display these patterns.



Video Segment (approximate times: 8:30-8:49 and 10:03-12:21): You can find the first part of this segment on the session video approximately 8 minutes and 30 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing. The second part of this segment begins approximately 10 minutes and 3 seconds after the Annenberg/CPB logo.

In the video segment, Professor Kader and participants examine the distribution of a data set to help explain an intriguing aspect of the variation they find. Professor Kader also introduces the uses of graphical and tabular representations. **[See Note 2]**

In Problem H2 of Session 1, you looked at the mint marks of 100 Jefferson nickels. Then you answered a series of questions about those coins. In this video segment, the onscreen participants examine the distribution of coins to come up with a hypothesis about where the coins without mint marks (N) were minted. After you watch the video segment, reflect on the questions below:

Write and Reflect

Problem A1. Did you arrive at the same conclusions about where the coins were minted as the onscreen participants did? Was your reasoning similar or different from theirs?

Note 2. If you're working with a group, you may want to organize a coin activity based on Problem H2 from Session 1 before watching the video segment. Here's how the activity can be organized:

Label the corners of a large, square poster board with the letters P (for Philadelphia), D (for Denver), S (for San Francisco), and N (for none). Separate 100 nickels according to mint mark (using magnifying glasses, if needed), and place them on the corner of the poster board corresponding to their mint location. Discuss the following question:

- Based on what you see, what can you say about the way the coins are distributed among the four different mint marks?

Statisticians find it useful to think in fractional terms, i.e., proportions or percentages. In order to start thinking about the coin data in fractional terms, cross two pieces of string to divide the poster board into four sections (one for each mint location). Adjusting the strings to keep the coins from each of the four mint marks separate, slide the coins to form a circle. This will result in a pie chart. Discuss the following questions:

1. Based on what you see in the pie chart, what can you say about the way the coins are distributed among the four different mint marks? Specifically, what fraction of the total would you guess is represented by each group?
2. Which location has the most coins? Why do you think this location is the most common?
3. Which location has the least coins? Why do you think this location is the least common?

Watch the video segment, and compare your hypotheses with those of the online participants.

Part B: Line Plots (40 min.)

Counting Raisins

Let's begin with a recap of Problem B7 from Session 1.

1. Ask a question:

How many raisins are in a half-ounce box of Brand X raisins? The weight of a box of raisins appears on the package, but the number of raisins in the box does not. In this activity, you will investigate how many raisins are in a box of a particular brand, which we will call Brand X.

2. Collect appropriate data:

For a non-interactive version of this activity, get some packages of half-ounce raisins and count them yourself. **[See Note 3]**

We opened and counted 17 boxes of raisins, which resulted in the following counts:

Number of Raisins in 17 Half-Ounce Boxes

29, 27, 27, 28, 31, 26, 28, 28, 30, 29, 26, 27, 29, 29, 25, 28, 28

Try It Online! www.learner.org

This problem can be explored online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 2, Part B, Problem B1.

Remember that data of this type is quantitative or numerical, since a raisin count is a measure of quantity. The raisin counts are also called discrete data, meaning that the measurements are obtained by counting and therefore must be whole numbers. We can also order the raisin counts, since it makes sense to say that one box has more or fewer raisins than another.

Problem B1. Using the data above, answer this question: How many raisins are in a half-ounce box of Brand X raisins? Answer the question in whatever manner seems the most descriptive.

You may have come up with a single number for the answer to the question above, or perhaps you came up with an interval of values for the answer. But because different boxes have different raisin counts, a single number will not provide a complete answer to the question. We cannot, for instance, say that a box contains 28 raisins—some do, but some don't. The raisin counts vary from box to box, so answers to this question must consider the variation in the data.

Problem B2. Suppose we count 17 half-ounce boxes of Brand Y raisins, and the resulting raisin counts are as follows:

Number of Raisins in 17 Half-Ounce Boxes

28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28

Answer this statistical question: How many raisins are in a half-ounce box of Brand Y raisins?

Problem B2 suggests that when there is no variation in the data, it is very easy to answer a statistical question about it.

Note 3. If you are doing this activity hands-on, count and record the number of raisins in each of your 17 boxes. Then consider the following question:

- For your brand, do all the boxes have the same number of raisins? Would you say that variation is present in your data?

Part B, cont'd.

Problem B3. Does Problem B2's data strongly suggest that the next box will have 28 raisins? Does it prove that the next box will have 28 raisins? If so, why? If not, would there be a way to prove, statistically, that the next box must have 28 raisins?

Problem B4. Now go back to the question in Problem B1, and use that data to describe the raisin count in a box of Brand X raisins. This time, try to consider the variation in the number of raisins per box.

Making a Line Plot

As we mentioned before, looking at quantitative data—numbers that come from measurements—provides answers to statistical questions. We are mainly concerned with situations where the measurements differ; that is, where there is variation in the data. Our answers to statistical questions must take this variation into account, so we need appropriate tools for describing the differences in measurements. **[See Note 4]**

One such tool is a graphical representation known as a line plot. In a line plot, we mark each possible value between the minimum and maximum data values and then stack dots above each of these values to represent actual counts. A line plot is sometimes called a dot plot.

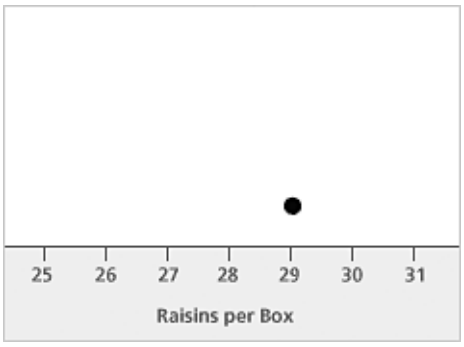
Recall the raisin counts for 17 boxes of Brand X raisins:

Number of Raisins in 17 Half-Ounce Boxes

29, 27, 27, 28, 31, 26, 28, 28, 30, 29, 26, 27, 29, 29, 25, 28, 28

To construct a line plot, we'll begin by setting up the horizontal axis for this set of data. Since the lowest (minimum) value is 26 and the highest (maximum) value is 31, we'll display this segment of the number line along the horizontal axis.

Next, for each raisin count, we'll place a dot above its corresponding value on the horizontal axis. For example, to display the count of our first box of Brand X raisins, we put a dot above the number 29.



To complete the line plot, we'll place a dot over the value 27, follow that with another dot over the value 27, and so forth, until there is a dot for each value in the data set.

Note 4. The goal of this section is to investigate a graphical representation that can help you better understand the variation in your data and provide various answers to this question: How many raisins are in a box of Brand X raisins?

Part B, cont'd.

Try It Online! **www.learner.org**

This problem can be explored online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 2, Part B, Problem B5.

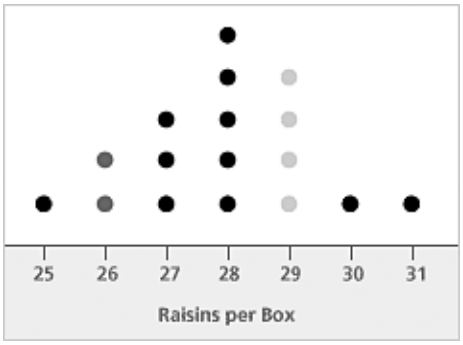
For a non-interactive version of this activity, use a piece of paper to add the rest of the data to the line plot we began above.

When the line plot is complete, the number of dots above each value indicates the frequency, or the number of times, that this particular raisin count appears in the data.

Problem B5. Does it make a difference that there is at least one of each discrete value between the lowest and highest values in the data (i.e., every raisin count between 25 and 31 has at least one box)?

Interpreting a Line Plot

Here is the line plot for the 17 raisin counts of Brand X raisins: **[See Note 5]**



Let's take a closer look at this graph. The horizontal axis corresponds to the number of raisins in a box. Each dot indicates one box of raisins, and the dots are placed above the numbers that indicate how many raisins are in the box. For instance, the four light gray dots tell us that four boxes contain 29 raisins. The two dark gray dots tell us that two boxes contain 26 raisins.

It is important to note that the raisin counts are ordered on the horizontal axis. A proper interpretation of this graph depends on this ordering.

Note 5. The line plot provides a picture of the distribution of the raisin counts. It shows us what values the raisin-count variable takes and how often each value occurs. It also shows the pattern of variation in the data.

If you are working with real raisins, construct a line plot for your data.

If you are working in groups, small groups can present their line plots to the whole group at this time.

Next, you will consider 10 questions about the distribution counts. Note that these questions all deal with the values of the counts or how often they occur. For example, Problem B6 (c)-(j) are all concerned with how many times a specific count occurs or how many times the counts occur within a specified range. Problem B8 is concerned with the principle that statistical problem solving requires answering questions in the presence of variation.

If you are working in groups, you can now return to your small groups to answer the questions in Problem B6.

Part B, cont'd.

Problem B6. Use the line plot to answer the following questions:

a. What is the minimum (smallest) raisin count for a box of Brand X raisins?

b. What is the maximum (largest) raisin count for a box?

c. How many boxes have between 26 and 28 raisins, inclusively (i.e., including 26 and 28)?

d. How many boxes have between 25 and 31 raisins, inclusively (i.e., including 25 and 31)?

e. Which raisin count occurred most frequently?

f. How many boxes contain more than 29 raisins?

g. How many boxes contain 29 or fewer raisins?

h. How many boxes contain fewer than 26 raisins?

i. How many boxes contain 25 or fewer raisins?

j. How many boxes contain between 26 and 29 raisins, inclusively?

Problem B7. Look back at the answers you gave in Problem B6 (f) and (g). Are these answers related? If so, how? And why? What about your answers to Problem B6 (h) and (i)?

Problem B8. Based on your observations above, give three descriptive statements that provide an answer to the question "How many raisins are there in a half-ounce box of Brand X raisins?" At least two of your statements should take into account the variation in the data.

Video Segment (approximate times: 12:50-13:50 and 14:42-15:33): You can find the first part of this segment on the session video approximately 12 minutes and 50 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing. The second part of this segment begins approximately 14 minutes and 42 seconds after the Annenberg/CPB logo.

In this video segment, participants attempt to answer the question "How many raisins are there in a box of Brand X raisins?" by collecting data, analyzing data, and interpreting the results. Watch the segment after you have completed Problems B1-B8, and compare your strategy with the onscreen participants'.

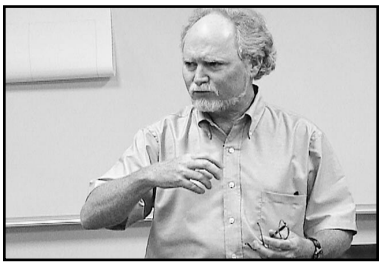
Is there a lot of overall variation in the data collected by Georgina's group? Are there places where there isn't a lot of variation in this data?

Part B, cont'd.

Intervals

When there is variation in data, there are many different answers to a statistical question, as your answer must take this variation into account. Frequently, answers to statistical questions are given in the form of intervals—ranges of values for data. Here are two common ways to use intervals to answer statistical questions:

1. Naming the interval in which all the data are located; that is, from the minimum data value (Min) to the maximum data value (Max). For example, in the Brand X raisin-count data, the interval is 25 to 31.
2. Naming an interval with the highest concentration of data; that is, an interval with little variation that contains a lot of data. For example, in the raisin-count data, a large proportion (14/17) of the Brand X raisin counts are between 26 and 29 (inclusively); this interval is 26 to 29.



Video Segment (approximate times 17:08-19:14): You can find this segment on the session video, approximately 17 minutes and 8 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing.

In this video segment, Professor Kader leads a discussion about two potential answers to the question "How many raisins are there in a box?"

Consider Paul and Phil's answers to Professor Kader's question. When might Paul's answer be more useful? How about Phil's? Which answer provides a better overall way of looking at the data? Why?

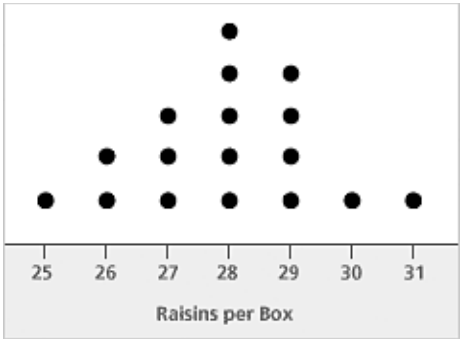
Sometimes it's useful to answer a statistical question with a single value that you've chosen to represent all of your data. The most frequently occurring value (the mode) may be a good choice. For example, in the Brand X raisin-count data, the most common raisin count is 28, which occurred five times. As we continue, we will encounter two other such representative values, the mean (the arithmetic average of the data set) and the median (the value in the exact center of an ordered list of data).

Part C: Frequency Tables (40 min.)

Making a Table

As you saw in Part A, a line plot is a graphical representation of data. For the raisin-count data, it showed how many times each raisin count occurred among the 17 boxes of Brand X raisins. You can also describe the same data using a frequency table, which shows the number of times each value occurs. The frequency table contains the same information as the line plot, but in tabular rather than graphical form. **[See Note 6]**

Problem C1. Use the line plot to complete the frequency table for the Brand X raisin counts. The first column lists each of the values that occurred in the raisin counts. The corresponding cell in the second column indicates the frequency—the number of times that that value occurred. For instance, only one box contained 25 raisins, so the frequency of 25 is 1.



Raisin Count	Frequency
25	1
26	
27	
28	
29	
30	
31	

Note 6. The frequency table that corresponds to the line plot contains the same information, but it's in another form (or representation). Part of the lesson here is to understand that there can be more than one way to represent data and to see the connection between two representations. You will be asked to consider the same 10 questions as before. Though you already know the answers, by solving the problems again, you will learn how to use the table, and you can compare the experience of using the table with that of using the line plot.

Keep in mind that different people see ideas in different ways. Some prefer graphical representations, and some prefer tabular representations. Luckily, in statistics we use both.

Part C, cont'd.

Problem C2. Use the frequency table to answer the following questions:

a. What is the minimum (smallest) raisin count for a box of Brand X raisins?

b. What is the maximum (largest) raisin count for a box?

c. How many boxes have between 26 and 28 raisins, inclusively (i.e., including 26 and 28)?

d. How many boxes have between 25 and 31 raisins, inclusively (i.e., including 25 and 31)?

e. Which raisin count occurred most frequently?

f. How many boxes contain more than 29 raisins?

g. How many boxes contain 29 or fewer raisins?

h. How many boxes contain fewer than 26 raisins?

i. How many boxes contain 25 or fewer raisins?

j. How many boxes contain between 26 and 29 raisins, inclusively?

Problem C3. Which questions in Problem C2 were easier to answer with a frequency table than with a line plot? Which were harder?

Cumulative Frequencies

Let's look at another useful way to describe variation in numerical data: A cumulative frequency specifies how many data values are of a particular number or smaller. **[See Note 7]**

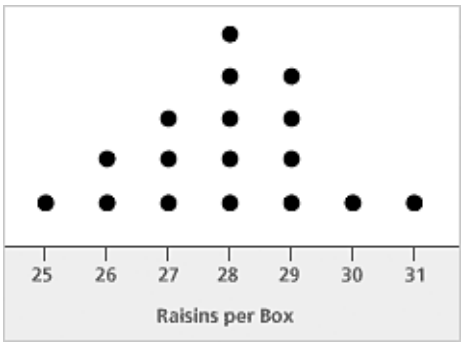
To obtain cumulative frequencies, it is first useful to obtain an ordered list of the data. Let's do this now with our original Brand X raisin data.

You may recall that the original listing of the raisin counts was not in order:

Number of Raisins in 17 Half-Ounce Boxes

29, 27, 27, 28, 31, 26, 28, 28, 30, 29, 26, 27, 29, 29, 25, 28, 28

We can obtain an ordered list from the line plot we created:



Note 7. The frequency table gave us the number of times a specific value occurred. We have also seen how an interval is used to describe how many raisins are in a box; for example, most of the boxes (14/17) contain between 26 and 29 raisins. We were able to determine how many of the counts fall in the interval 26-29 by adding the individual frequencies in this range.

The cumulative frequency function simplifies this process and gives us a more convenient device for obtaining frequencies for an interval of data—the cumulative frequency function has already done the adding! Keep in mind that with large data sets this would be an even greater advantage. If you have the cumulative frequencies, then the computation of the frequency within an interval is simply the difference between two numbers. This idea may not be obvious at first, but you'll see as you get to play with it a bit in this part of the session.

If you are working with actual raisins, make frequency and cumulative frequency tables with your own data.

Part C, cont'd.

- The smallest raisin count is 25. Therefore, the ordered list begins with 25. As there is only one box of count 25, we look to the next count to find the next number in the sequence.
- The next-smallest raisin count is 26. There are two boxes of size 26. The ordered list is now 25, 26, 26.
- The next-smallest raisin count is 27. There are three boxes of count 27. The ordered list is now 25, 26, 26, 27, 27, 27.

This table shows the final ordered list of Brand X raisin counts:

Position	Raisin Count
1	25
2	26
3	26
4	27
5	27
6	27
7	28
8	28
9	28
10	28
11	28
12	29
13	29
14	29
15	29
16	30
17	31

In this problem, the cumulative frequency specifies how many boxes have raisin counts *of a particular count or smaller*. Reading this table in terms of cumulative frequency tells us, for example, that there are 11 values that are 28 or smaller and 15 values that are 29 or smaller.

Part C, cont'd.

A cumulative frequency table lists the cumulative frequency for each value in the data set. To construct a cumulative frequency table, start with the smallest raisin count in the data. According to the ordered list, there is only one box with 25 raisins or fewer, so we record this in the cumulative frequency column. Moving on to the next count in the ordered list, we see that there are three boxes with 26 or fewer raisins.

Raisin Count	Cumulative Frequency
25	1
26	3
27	
28	
29	
30	
31	

Problem C4. Use the ordered list of raisin counts given earlier to complete the cumulative frequency table above.

Problem C5. Use the cumulative frequency table to answer the following questions:

a. What is the minimum (smallest) raisin count for a box of Brand X raisins?

b. What is the maximum (largest) raisin count for a box?

c. How many boxes have between 26 and 28 raisins, inclusively (i.e., including 26 and 28)?

d. How many boxes have between 25 and 31 raisins, inclusively (i.e., including 25 and 31)?

e. Which raisin count occurred most frequently?

f. How many boxes contain more than 29 raisins?

g. How many boxes contain 29 or fewer raisins?

h. How many boxes contain fewer than 26 raisins?

i. How many boxes contain 25 or fewer raisins?

j. How many boxes contain between 26 and 29 raisins, inclusively?

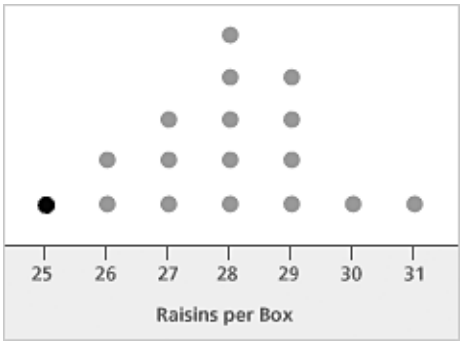
Problem C6. Which of the questions in Problem C5 were easier to answer with a cumulative frequency table? Which were more difficult?

The cumulative frequency table becomes more important in data sets with a wide spread of values. For example, while it may not be that useful to know that 1.7% of students scored exactly 510 on a standardized test, it may be much more useful to know that 53.6% of students scored no higher than 510 on the same test. In this way, a cumulative frequency table can be used to calculate percentiles.

Part C, cont'd.

Another Method

Here is another way that you could use either the line plot or the frequency table to obtain the cumulative frequencies. Look at the number of boxes with 25 or fewer raisins, which are highlighted in black on the line plot and in bold on the frequency table:



Raisin Count	Frequency
25	**1**
26	2
27	3
28	5
29	4
30	1
31	1

There is one box with 25 or fewer raisins.

Now look at the number of boxes with 26 or fewer raisins, highlighted in gray on the line plot and in bold on the frequency table:

Raisin Count	Frequency
25	**1**
26	**2**
27	3
28	5
29	4
30	1
31	1

There are three boxes with 26 or fewer raisins.

Problem C7. You can add a third column to the table to indicate cumulative frequencies. Use the line plot or the frequency table to complete the cumulative frequency values.

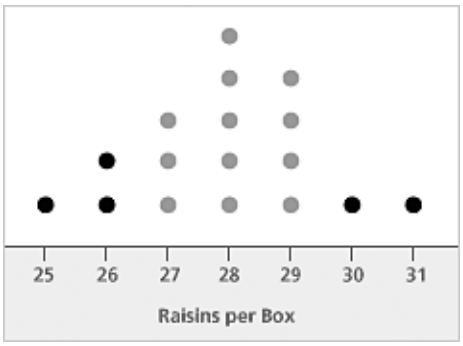
Raisin Count	Frequency	Cumulative Frequency
25	1	1
26	2	3
27	3	
28	5	
29	4	
30	1	
31	1	

[See Tip C7, page 54]

Part C, cont'd.

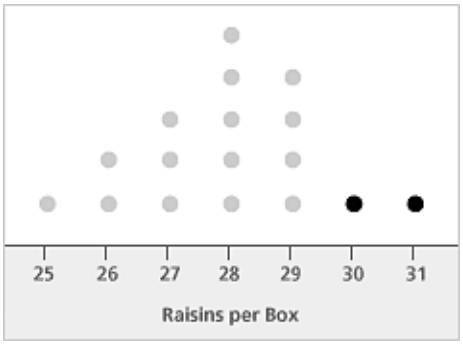
Intervals and Ranges

We use cumulative frequencies to describe intervals and ranges of data. For example, consider the boxes with between 27 and 29 raisins, inclusively, which are represented in gray on the line plot:



The number of boxes with between 27 and 29 raisins (12) is easy to determine from this line plot, but in problems with very large data sets, this might not be the case.

Here's how a cumulative frequency table can be used to answer the question of how many boxes have between 27 and 29 raisins, inclusively. First, look at the number of boxes with counts of 29 or smaller. There are 15 of these, represented in light gray on the line plot and in bold on the frequency table:



Raisin Count	Frequency	Cumulative Frequency
25	**1**	1
26	**2**	3
27	**3**	6
28	**5**	11
29	**4**	**15**
30	1	16
31	1	17

Part C, cont'd.

Remove the boxes that have fewer than 27 raisins. There are three of these, highlighted in dark gray on the line plot and in bold italics on the frequency table:

Size	Frequency	Cumulative Frequency
25	***1***	1
26	***2***	***3***
27	**3**	6
28	**5**	11
29	**4**	**15**
30	1	16
31	1	17

The number of remaining boxes is:

15 - ***3*** = 12

Therefore, there are 12 boxes that contain between 27 and 29 raisins.

Problem C8. Use the method described above to find the following:

a. The number of boxes that contain between 26 and 30 raisins, inclusively

b. The number of boxes that contain between 27 and 31 raisins, inclusively

c. The number of boxes that contain more than 28 raisins

Part D: The Median (25 min.)

From Ordered Lists and Line Plots

A common way to summarize data is to use numerical summaries, many of which are based on the ordered data. For example, the largest and smallest data values (minimum and maximum) are the first and last values in the ordered data. If we know the first and last values in an ordered list, we know that all the data values are between these two numbers.

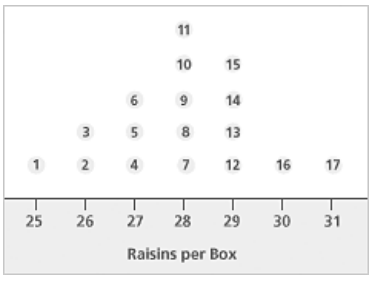
Another numerical summary that is based on ordered data is the median, which is the middle value in an ordered list. Let's find and interpret the median, using our raisin data. **[See Note 8]**

We'll begin by examining the ordered list of Brand X raisin counts:

Position	Raisin Count
1	25
2	26
3	26
4	27
5	27
6	27
7	28
8	28
9	28
10	28
11	28
12	29
13	29
14	29
15	29
16	30
17	31

Note 8. This session provides a quick look at the median, which will be explored in more detail in Session 4.

Three different representations are used in Part D. The median is first examined in the ordered list. It is then determined by looking at the ordering in the line plot:



Finally, cumulative frequencies are used to determine the median.

If you are working with actual raisins, find the median for your data using each of the three representations: ordered list, line plot, and cumulative frequencies.

Part D, cont'd.

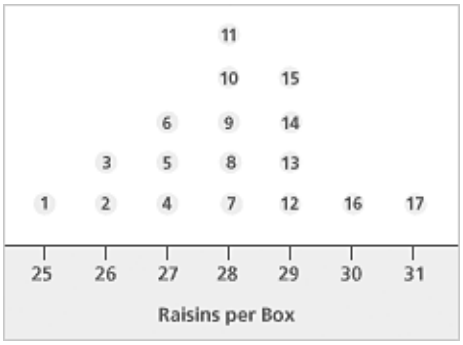
Problem D1. Which position corresponds to the median (the position in the middle of the list)? How many raisins are there in the box at this position? **[See Tip D1, page 54]**

The median is the value in the exact center of a data set—in other words, there are as many values above it as there are below it. In this case, the median is in the ninth position, since there are eight values below it and eight above. Note that in any data set with 17 values, the ninth value in the ordered list will always be the median.

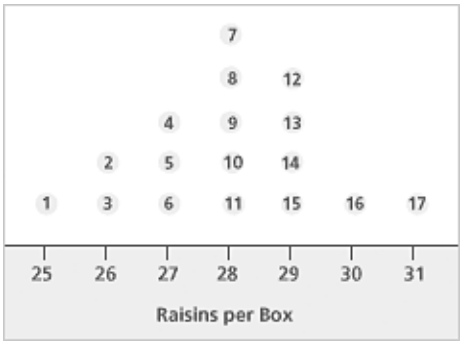
Problem D2. Suppose that the data were ordered from highest to lowest instead of from lowest to highest. How would you find the median then?

We can use the median along with the minimum and maximum to describe variation in data. The median divides the raisin-count data into two groups: the data values below the median and the data values above the median. Note that each group has eight data values, which is approximately half the data. Consequently, approximately half of the raisin counts are in the interval 25 to 28 (from the minimum to the median), and approximately half of the raisin counts are in the interval 28 to 31 (from the median to the maximum).

We can also determine the median by looking at a line plot. For the line plot of the raisin counts, you can identify the 17 positions in the ordered data as follows:



Alternatively, you could number the 17 positions in this way:



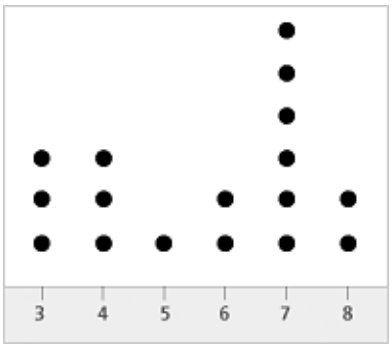
These two line plots are identical, since there is no need to distinguish the order of raisin boxes that have the same number of raisins.

Again we see that the ordered lists contain 17 data values, and there are 17 positions in the ordered list. Position 9 contains the middle value because eight positions precede Position 9, and eight positions follow Position 9. Position 9 corresponds to a box that contains 28 raisins, so 28 is the median. This can be determined from either of the line plots above.

Part D, cont'd.

Problem D3. Find the median of this data set: 72, 68, 63, 70, 84, 75, 72, 70, 82.

Problem D4. Find the median of the data set for this line plot:



From Cumulative Frequency Tables

You can also use a cumulative frequency table to find the median of the raisin count:

Raisin Count	Cumulative Frequency
25	1
26	3
27	6
28	11
29	15
30	16
31	17

The first position contains raisin count 25, the third position contains count 26, and so on.

Problem D5. According to the cumulative frequency table, what count is in the second position?

Problem D6. The sixth position corresponds to a box containing 27 raisins. How many raisins are in the boxes in the fourth and fifth positions? **[See Tip D6, page 54]**

Problem D7. How would you use the cumulative frequency table to find the median?

[See Tip D7, page 54]

Part E: Bar Graphs and Relative Frequencies (30 min.)

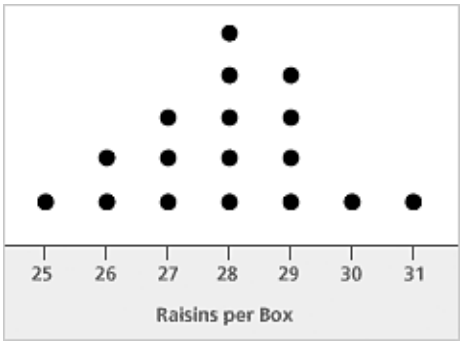
Frequency Bar Graphs

The line plot is a useful graph for examining small sets of data. It's especially helpful as a device for learning basic statistical ideas. But for larger data sets, it can be awkward to create, since for each data value there is a corresponding dot. That's a lot of dots for data sets with hundreds or thousands of values! You can, however, replace a line plot with a frequency bar graph.

Let's look at the transition from line plot to frequency bar graph.

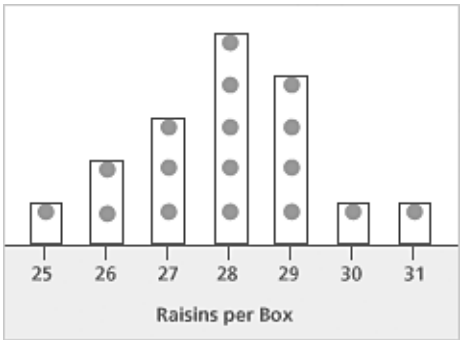
We'll start with the line plot we've been using. Remember that the number of dots over each value on the horizontal axis corresponds to the frequency of that data value:

Beginning Stage: Line Plot



Now draw a rectangle over each value, with a height corresponding to the frequency of that value:

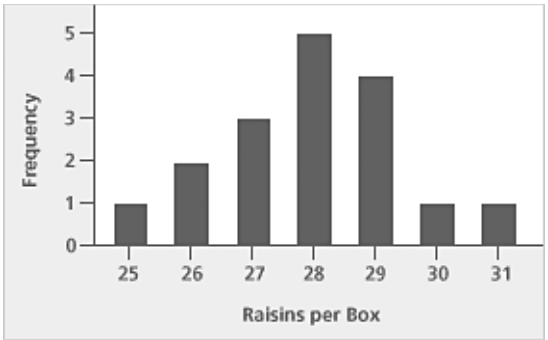
Intermediate Stage: Dots and Rectangles



Part E, cont'd.

Now remove the dots, and add a vertical scale that indicates the frequency of each value on the horizontal scale:

Final Stage: Frequency Bar Graph



The frequency bar graph contains the same information as the line plot for the counts of raisin boxes, but it doesn't indicate the raisin count for each individual box. The height of each bar or rectangle tells us the frequency for the corresponding raisin count.

Relative Frequency

Although the frequency bar graph is useful in many ways, it, like the line plot, can be an awkward graph for large data sets, since the vertical axis corresponds to the frequency of each data value. For large data sets, some data values occur many times and have a high frequency. Consequently, the vertical axis would have to be scaled according to the largest frequency. Imagine the sheet of paper you'd need for the economy-size box of raisins!

An alternative is to use relative frequency, or frequency as a proportion of the whole set. A relative or proportional comparison is usually more useful than a comparison of absolute frequencies. For example, the statement "Five of the 17 boxes have 28 raisins" is more useful than the statement "Five boxes have 28 raisins."

In this case, the relative frequency of the count 5 is 5/17, which can also be written in decimal form as .294 (rounded to three digits). To find the percentage, multiply the decimal by 100 to obtain 29.4%. This means that 29.4% of the raisin boxes contain 28 raisins.

Here is a frequency table for the raisin count, with the corresponding relative frequencies written as fractions, decimals, and percentages:

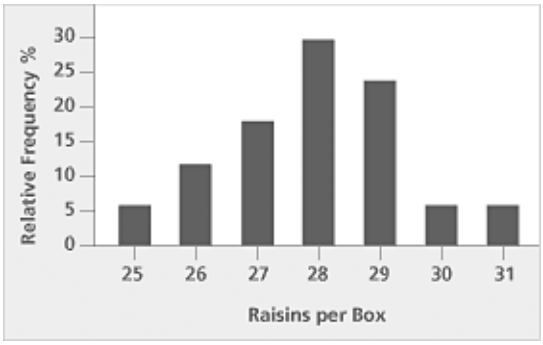
Raisin Count	Frequency	Relative Frequency		
		Fraction	Decimal	Percentage
25	1	1/17	.059	5.9
26	2	2/17	.118	
27	3			17.6
28	5			
29	4		.235	
30	1	1/17		
31	1			

Part E, cont'd.

Problem E1. Complete the table on the previous page. Give decimals to three decimal places and percentages to the nearest tenth of a percent.

Notice that the relative frequencies expressed as fractions add up to 17/17, which equals 1. The relative frequencies expressed as decimals also sum to 1, and the relative frequencies expressed as percentages add up to 100%. The total of the relative frequencies expressed as decimals, however, may not always be exactly 1 because of round-off error; they will occasionally add to 1.002 or 0.997, for example, or something very close to 1. Accordingly, the total percentage may not sum to exactly 100%. To decrease round-off error, we would have to increase the number of decimal places used when rounding.

A relative frequency bar graph looks just like a frequency bar graph except that the units on the vertical axis are expressed as percentages. In the raisin example, the height of each bar is the relative frequency of the corresponding raisin count, expressed as a percentage: **[See Note 9]**



One advantage to using relative frequencies is that the total of all relative frequencies in a data set should be 1 (or very close to 1, depending on round-off error), or 100%. In this way, a relative frequency bar graph allows you to think of the data in terms of the whole set in contrast to a frequency bar graph, which only provides you with individual counts.

Comparing Representations

In statistics, it can be difficult to provide a specific answer to a question because of the variation present in the data. Statistical analysis allows us to organize data in different ways so that we can draw out potential patterns in the variation and give better answers to the questions posed.

Try It Online! **www.learner.org**

This problem can be explored online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 2, Part E.

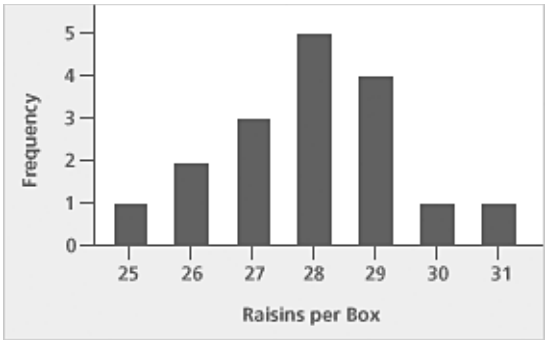
For a non-interactive version of this activity, review and compare the following representations of data we have explored, both graphical and tabular.

Note 9. The next transition in the representation is to replace frequencies with relative frequencies. The relative frequency bar graph looks exactly the same as the frequency bar graph. It is important to note that the sizes of the bars remain the same whether you use frequencies or relative frequencies.

If you are working with actual raisins, draw a frequency bar graph and a relative frequency bar graph with your data. Do your graphs look the same?

If you are working in groups, small groups can present their bar graphs to the whole group at this time.

Part E, cont'd.



Raisin Count	Frequency	Relative Frequency		
		Fraction	Decimal	Percentage
25	1	1/17	.05882	5.882
26	2	2/17	.11764	11.764
27	3	3/17	.17647	17.647
28	5	3/17	.29411	29.411
29	4	4/17	.23529	23.529
30	1	1/17	.05882	5.882
31	1	1/17	.05882	5.882

Raisin Count	Frequency	Cumulative Frequency
25	1	1
26	2	3
27	3	6
28	5	11
29	4	15
30	1	16
31	1	17

It's important to note that several kinds of answers can be given when there is variation in your data. Some answers may be stated as intervals, and some answers, like the mode and the median, use a specific value to represent all the different data values.

Video Segment (approximate times: 23:03-23:45 and 24:48-25:30): You can find the first part of this segment on the session video approximately 23 minutes and 3 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing. The second part of this segment begins approximately 24 minutes and 48 seconds after the Annenberg/CPB logo.

In this video segment, meteorologist Kim Martucci demonstrates how she solves the statistical problem of predicting the weather. Watch this segment after you have completed Session 2. What are Kim Martucci's strategies for predicting the weather? How are they similar to your strategies for counting raisins? How are they different?

Homework

Problem H1. For the following data sets, create line plots, frequency tables, and cumulative frequency tables. Use your results to answer the question "How many raisins are in a half-ounce box of raisins?" for each brand.

Brand A

23	25	25	26	26	26	26	27	27	27	27
28	29	29	29	30	30	31	31	31	32	32
32	33	34	34	35	35	36	39			

Brand B

17	22	24	24	25	25	25	25	26	26	26
26	26	26	27	27	27	27	28	29	29	29
29	29	29	30	30						

Brand C

25	25	25	26	26	26	26	26	27	27	27
28	28	28	28	28	28	28	28	28	29	29
29	30	30	31	32	32					

Brand D

23	24	25	25	25	27	27	27	27	27	27
27	27	28	28	29	29	29	29	29	29	30
31	32	32	33	33	33	34	34	35	35	35
36	36	38								

Problem H2. Based on your analyses in Problem H1, which brand of raisins would you buy? Explain.

Problem H3.

a. Use the representations of the data you developed for Problem H1 to determine the minimum and maximum raisin counts and the median raisin count for each brand of raisins (A, B, C, and D).

b. Which brand has the most variation? Which has the least variation?

c. Which brand typically has more raisins? Which brand typically has fewer raisins?

d. Does your work on this problem change your answer to Problem H2?

Problem H4. Choose one of the brands listed above and create a relative frequency table and relative frequency bar graph for it. (The solution for Brand A is given.)

Take It Further

Problem H5. Create two data sets that have the same mean, the same median, and the same mode, but are not identical data sets. How could you distinguish these sets from each other?

Homework, cont'd.

Suggested Readings

These readings are available as downloadable PDF files on the *Data Analysis, Statistics, and Probability* Web site. Go to:

www.learner.org/learningmath

Friel, Susan, Bright, George, and Curcio, Frances (November-December, 1997). "Understanding Students' Understanding of Graphs," *Mathematics Teaching in the Middle School, 3* (3), 224-227.

Kader, Gary, and Perry, Mike (November-December, 1997). "Pennies From Heaven—Nickels From Where?," *Mathematics Teaching in the Middle School, 3* (3), 240-248.

Putt, Ian; Jones, Graham; Thornton, Carol; Langrall, Cynthia; Mooney, Edward; and Perry, Bob (Autumn, 1999). "Young Students' Informal Statistical Knowledge," *Teaching Statistics, 21* (3), 74-78.

Tips

Part C: Frequency Tables

Tip C7. A quicker way to find the cumulative frequency is to add the previous cumulative frequency to the frequency of the new value. So for 28 or fewer, add the previous cumulative frequency (6) to the frequency of the new value (5) to get the new cumulative frequency (11).

Part D: The Median

Tip D1. One way to find the median is to continue to remove the highest and lowest values in the data set until only the median remains.

Tip D6. How many raisins are in the third position? Why can't this be the same as the number of raisins in the fourth position?

Tip D7. First find the position in which the median is located (see Problem D1). Then look for the corresponding number on the table.

Solutions

Part B: Line Plots

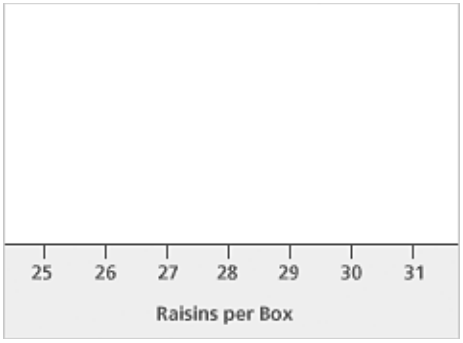
Problem B1. There are many possible answers. For example, we can be fairly confident that a box of Brand X raisins will have between 25 and 31 raisins.

Problem B2. We can be almost certain that a new box of Brand Y raisins will have 28 raisins.

Problem B3. Yes, the data strongly suggests that the next box will have 28 raisins, but it does not prove this. We would need to be sure that *all* boxes would have 28 raisins, not just our sample of 17 boxes. Statistically, there is no way to guarantee that the next box must have 28 raisins without sampling all boxes or without knowing something about the process involved in making Brand Y raisins. This is true regardless of how many boxes we sample; it only becomes more and more likely that the next box will have 28 raisins.

Problem B4. We can still be fairly confident that a box of Brand X raisins will have between 25 and 31 raisins. In looking at the variation, though, it is quite likely that a box of Brand X raisins will have between 27 and 29 raisins. Still, no single number can describe our expectation for the number of raisins in a box of Brand X raisins.

Problem B5. The range of possible data values is between 25 and 31 raisins, so the horizontal axis will include each number from 25 to 31:



It would not matter if, for example, there were no values of 30 in the data set. All values in the range of possible data values must be included in the line plot, just as all numbers are indicated on a number line even though some of them may not be included in a list.

Problem B6

a. The minimum raisin count is the leftmost dot, which indicates 25 raisins.

b. The maximum raisin count is the rightmost dot, which indicates 31 raisins.

c. Counting the dots tells us that a total of 10 boxes contains between 26 and 28 raisins.

d. All 17 boxes do.

e. The most frequent count is the tallest stack of dots, 28 raisins. There are five boxes with this frequency.

f. Two boxes contain more than 29 raisins.

g. Fifteen boxes contain 29 or fewer raisins.

h. One box contains fewer than 26 raisins.

i. One box contains 25 or fewer raisins.

j. Fourteen boxes contain between 26 and 29 raisins.

Solutions, cont'd.

Problem B7. For Problem B6 (f) and (g), if you already know how many boxes contain more than 29 raisins, all other boxes must contain 29 or fewer raisins. Instead of counting a large number of boxes for Problem B6 (g), you could subtract the answer to Problem B6 (f), which is 2, from the total number of boxes, which is 17, to get your answer, 15.

As for Problem B6 (h) and (i), they are identical questions because the data is discrete. There is no way to have between 25 and 26 raisins, so asking how many boxes have fewer than 26 raisins is the same as asking how many have 25 or fewer.

Problem B8

a. The box of raisins should have between 25 and 31 raisins.

b. The box of raisins is very likely to have between 26 and 29 raisins.

c. The box of raisins is unlikely to have 28 raisins, but this is the most likely number from our sample.

Part C: Frequency Tables

Problem C1. Here's what the completed frequency table should look like:

Raisin Count	Frequency
25	1
26	2
27	3
28	5
29	4
30	1
31	1

Problem C2.

a. The minimum raisin count is 25 raisins.

b. The maximum raisin count is 31 raisins.

c. Adding the frequencies results in a total of 10 boxes containing between 26 and 28 raisins.

d. All 17 boxes do.

e. The count of 28 raisins has the highest frequency, which is five.

f. Two boxes contain more than 29 raisins: one has 30, and one has 31.

g. Fifteen boxes contain 29 or fewer raisins.

h. One box contains fewer than 26 raisins.

i. One box contains 25 or fewer raisins.

j. Fourteen boxes contain between 26 and 29 raisins.

Solutions, cont'd.

Problem C3. Answers vary. Questions about individual raisin counts tend to be easier to answer with a frequency table, as there is no counting required, but questions about ranges of values are often easier with a line plot.

Problem C4. Here's what the completed table should look like:

Raisin Count	Cumulative Frequency
25	1
26	3
27	6
28	11
29	15
30	16
31	17

Problem C5. The answers are identical to those in Problem C2.

Problem C6. Questions about ranges may be easier with the cumulative frequency table. C5 (d) and (g) in particular are easier, since their answers can be read directly from the table. Questions about individual frequencies are more difficult because they require subtraction to go from the cumulative frequency table to the frequency table.

Problem C7. Here's what the completed table should look like:

Raisin Count	Frequency	Cumulative Frequency
25	1	1
26	2	3
27	3	6
28	5	11
29	4	15
30	1	16
31	1	17

Problem C8.

a. The cumulative frequency for 30 raisins is 16, and the cumulative frequency for 25 raisins is 1, so the number of boxes containing between 26 and 30 raisins is 15. Note that we use the frequency for 25 raisins as the lower boundary if we want to include 26 in the count.

b. The cumulative frequency for 31 raisins is 17, and the cumulative frequency for 26 is 3, so the answer is 14 boxes.

c. The cumulative frequency for 28 raisins is 11, so all six other boxes (17 - 11) must have more than 28 raisins.

Solutions, cont'd.

Part D: The Median

Problem D1. The median is in the ninth position, with 28 raisins.

Problem D2. You would find the median in the same way, by finding the value that has an equal number of values above and below it. You could still do this by removing the highest and lowest values in the data set until only the median remains.

Problem D3. The ordered list is 63, 68, 70, 70, 72, 72, 75, 82, 84. The value in the center of this list is 72, which makes it the median.

Problem D4. The median is the ninth number in the ordered list, which, in this case, is 6.

Problem D5. The second value is 26, since its position (2) is higher than the cumulative frequency of 25 (1), but not higher than the cumulative frequency of 26 (3).

Problem D6. They also have 27 raisins.

Problem D7. We know that the median is in the ninth position (of 17 total boxes), which falls between the cumulative frequency of 27 (6) and 28 (11); therefore, the median is 28.

Part E: Bar Graphs and Relative Frequencies

Problem E1. Here's what the completed table should look like:

Raisin Count	Frequency	Relative Frequency		
		Fraction	Decimal	Percentage
25	1	1/17	.059	5.9
26	2	2/17	.118	11.8
27	3	3/17	.176	17.6
28	5	5/17	.294	29.4
29	4	4/17	.235	23.5
30	1	1/17	.059	5.9
31	1	1/17	.059	5.9

Solutions, cont'd.

Homework

Problem H1.

Brand A

Raisin Count	Frequency	Cumulative Frequency
23	1	1
24	0	1
25	2	3
26	4	7
27	4	11
28	1	12
29	3	15
30	2	17
31	3	20
32	3	23
33	1	24
34	2	26
35	2	28
36	1	29
37	0	29
38	0	29
39	1	30

There are between 23 and 39 raisins in a box. It is 67% likely (20 of 30) that a box will have between 26 and 32 raisins.

Brand B

Raisin Count	Frequency	Cumulative Frequency
17	1	1
18	0	1
19	0	1
20	0	1
21	0	1
22	1	2
23	0	2
24	2	4
25	4	8
26	6	14
27	4	18
28	1	19
29	6	25
30	2	27

There are between 17 and 30 raisins in a box. It is 78% likely (21 of 27) that a box will have between 25 and 29 raisins.

Solutions, cont'd.

Brand C

Raisin Count	Frequency	Cumulative Frequency
25	3	3
26	5	8
27	3	11
28	9	20
29	3	23
30	2	25
31	1	26
32	2	28

There are between 25 and 32 raisins in a box. It is 82% likely (23 of 28) that a box will have between 25 and 29 raisins.

Brand D

Raisin Count	Frequency	Cumulative Frequency
23	1	1
24	1	2
25	3	5
26	0	5
27	8	13
28	2	15
29	6	21
30	1	22
31	1	23
32	2	25
33	3	28
34	2	30
35	3	33
36	2	35
37	0	35
38	1	36

There are between 23 and 38 raisins in a box. It is 83% likely (30 of 36) that a box will have between 27 and 36 raisins

Problem H2. Answers will vary. Many will choose Brand A, since it has the greatest likelihood (50%) of having at least 30 raisins and has the largest median (29.5). Some will choose Brand C, since it is very unlikely to have 25 raisins or less in a box. Of course, those who don't like raisins will choose Brand B!

Solutions, cont'd.

Problem H3.

a. Brand A: minimum = 23, median = 29.5, maximum = 39

Brand B: minimum = 17, median = 26, maximum = 30

Brand C: minimum = 25, median = 28, maximum = 32

Brand D: minimum = 23, median = 29, maximum = 38

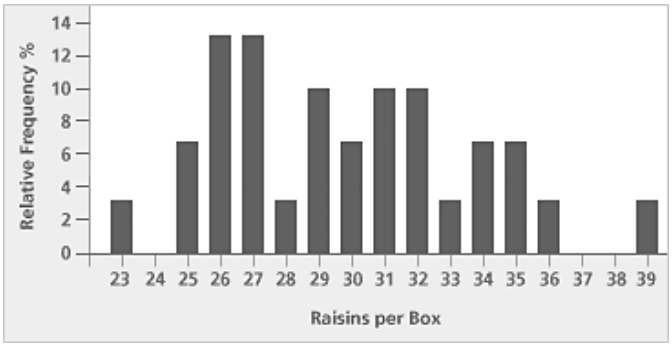
b. Brands A and D each have a larger range than B and C. Although Brand A has the wider range (39 – 23 = 16), Brand D has more extreme values than Brand A.

c. Brand A typically has the most raisins; it has the highest maximum and highest median. Brand B has the fewest raisins.

d. Answers will vary.

Problem H4. Here are the calculations if you had chosen Brand A:

Raisin Count	Frequency	Relative Frequency		
		Fraction	Decimal	Percentage
23	1	1/30	.033	3.3
24	0	0/30	.000	0.0
25	2	2/30	.067	6.7
26	4	4/30	.133	13.3
27	4	4/30	.133	13.3
28	1	1/30	.033	3.3
29	3	3/30	.100	10.0
30	2	2/30	.067	6.7
31	3	3/30	.100	10.0
32	3	3/30	.100	10.0
33	1	1/30	.033	3.3
34	2	2/30	.067	6.7
35	2	2/30	.067	6.7
36	1	1/30	.033	3.3
37	0	0/30	.000	0.0
38	0	0/30	.000	0.0
39	1	1/30	.033	3.3



Solutions, cont'd.

Problem H5. There are many possible answers. Here's one:

Set A: 10, 10, 20, 20, 20, 30, 30

Set B: 19, 19, 20, 20, 20, 21, 21

In each set, the mean, median, and mode are all 20. They are not identical; one distinguishing characteristic is the variation in the data. Set A has more variation, with four elements that are each 10 away from the mean. In Set B, all elements are within one of the mean.