Session 4

The Five-Number Summary

Key Terms for This Session

Previously Introduced

- interval
- median
- mode

New in This Session

- box plot
- Five-Number Summary
- interquartile range
- midrange
- quartiles
- summary measures
- Three-Number Summary
- Two-Number Summary

Introduction

Sessions 2 and 3 explored different ways to organize data in order to draw out patterns in the variation. In these sessions, we investigated a variety of approaches to organizing data in graphs and tables and examined some of the different ways we can answer statistical questions, based on these representations.

In this session, we will explore how dividing data into groups can give us yet another way to answer statistical questions. **[See Note 1]**

Learning Objectives

The goal of this session is to learn how dividing data into groups can help us provide other types of answers to statistical questions. In this session, you will learn how to do the following:

- Understand the median as the value that divides the ordered data into two groups, with the same number of data values (approximately one-half) in each group
- Understand quartiles as the values that divide the ordered data into four groups, with the same number of data values (approximately one-fourth) in each group
- Investigate the concept of quartiles using several different representations
- Summarize an entire data set with a Five-Number Summary
- Summarize an entire data set with a box plot

Note 1. This session relies on a hands-on activity using spaghetti noodles. Gather the necessary materials before beginning the session.

Part A: Min, Max, and the Two-Number Summary (20 min.)

The Data Set

When working with a large collection of data, it can be difficult to keep an accurate picture of your data in mind. One way to make it easier to work with large data sets is to reduce the entire data set to just a few summary measures (or numbers that describe significant characteristics of the data). In this session, you will learn how to determine summary measures from ordered data. For convenience, we'll be looking at small data sets, but these methods and interpretations apply to larger data sets as well.

For the following activities, you will need these materials:

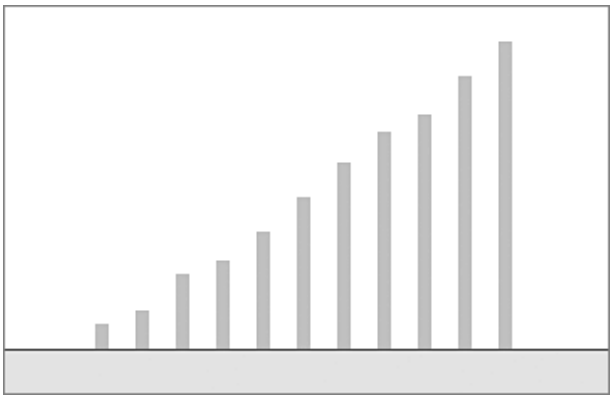
- a package of spaghetti or linguine
- a metric ruler with millimeter markings
- three pieces of paper or cardboard
- a pen or pencil

Ask a question:

How long is a broken piece of spaghetti?

Collect appropriate data:

Break several spaghetti noodles into pieces to obtain 11 noodles of varying lengths. Make sure that no two noodles in your set are the same length. Draw a horizontal line on a piece of paper or cardboard large enough to display all the noodles in a row. Next, arrange the 11 noodles in order from shortest to longest along the horizontal line. Your arrangement should look something like this:

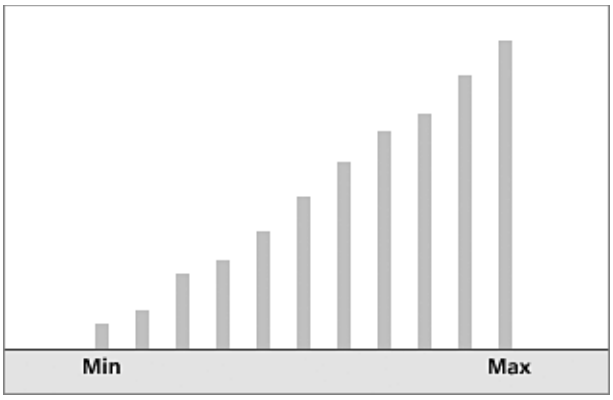


Part A, cont'd.

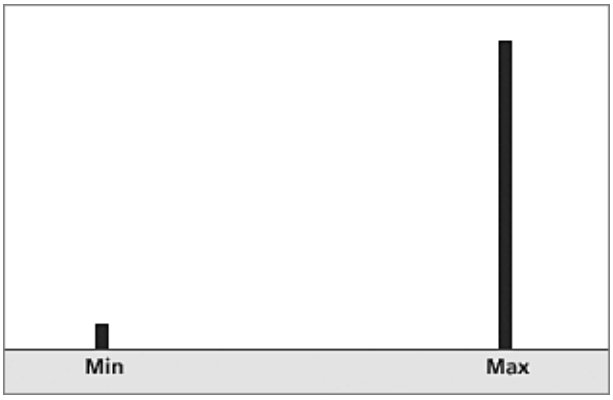
The Two-Noodle Summary

Two useful summary measures are the smallest (minimum) and largest (maximum) data values. To find these values in your ordered arrangement of noodles, remove all but the shortest and longest (keeping the others in size order for use later on).

Label the shortest "Min" (for minimum length) and label the longest "Max" (for maximum length):



We'll refer to these two noodles (Min and Max) as the "Two-Noodle Summary."



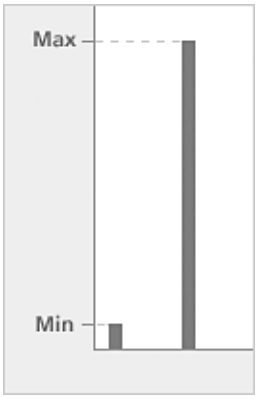
Problem A1. If you could see only Min and Max (as pictured at left), what could you say about any of the other nine noodles in the set?

Part A, cont'd.

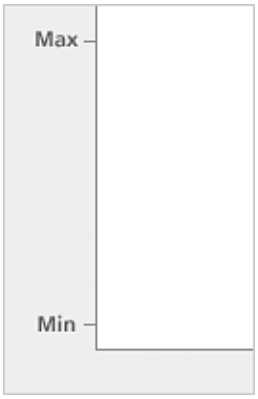
The Two-Number Summary

We will now determine the Two-Number Summary from the Two-Noodle Summary.

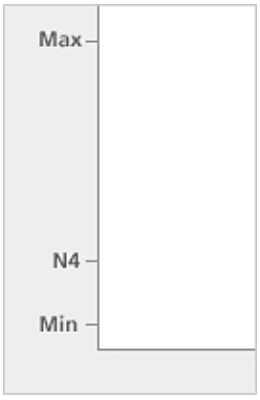
We will add a vertical axis and mark the lengths of the two noodles:



Remove the noodles. What remains is the Two-Number Summary.



If we recorded the length of the fourth noodle in the original set on the same vertical number line, it might look something like this:



Problem A2. What can you say about the length of noodle N4, given the information in the Two-Number Summary?

Problem A3. If you knew only the values of Max and Min, describe some information you would *not* know about the remaining nine noodles.

Problem A4. Suppose someone asked you to find the "typical" value of the noodle data in Problems A1-A3. How would you answer this question? How would you answer this question if you only had the information from the Two-Number Summary?

Video Segment (approximate time: 1:18-1:27): You can find this segment on the session video approximately 1 minute and 18 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing.

In this video segment, Professor Kader asks participants to identify the "typical" value in a data set. Watch this segment after completing Problem A4.

Note: The data set used by the onscreen participants is different from the one provided above.

How do participants define the "center" of a data set?

Part B: The Median and the Three-Number Summary (35 min.)

The Median

Another useful summary measure for a collection of data is the median. As you learned in Session 2, the median is the middle data value in an ordered list. Here's one way to find the median of our ordered noodles.

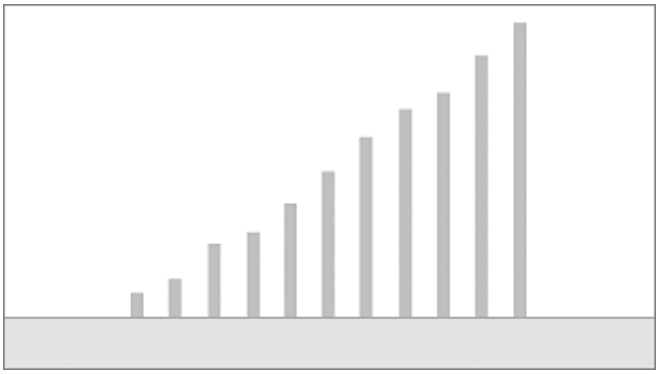
Try It Online! **www.learner.org**

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 4, Part B.

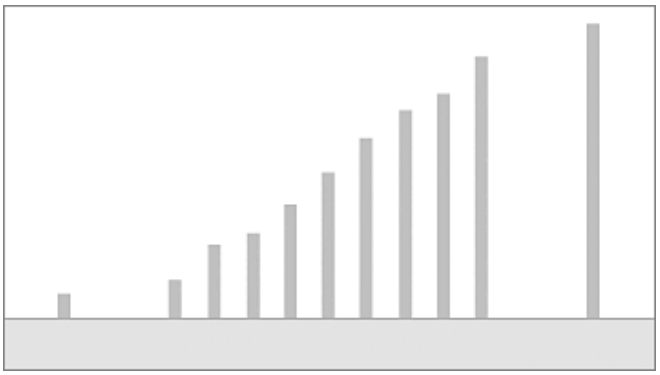
For a non-interactive version of this activity, look at the following illustrations.

We'll begin with the 11 noodles arranged in order from shortest to longest. We'll remove two noodles at a time, one from each end, and put them to the side. We'll continue this process until only one noodle remains. This noodle is the median, which we'll label "Med."

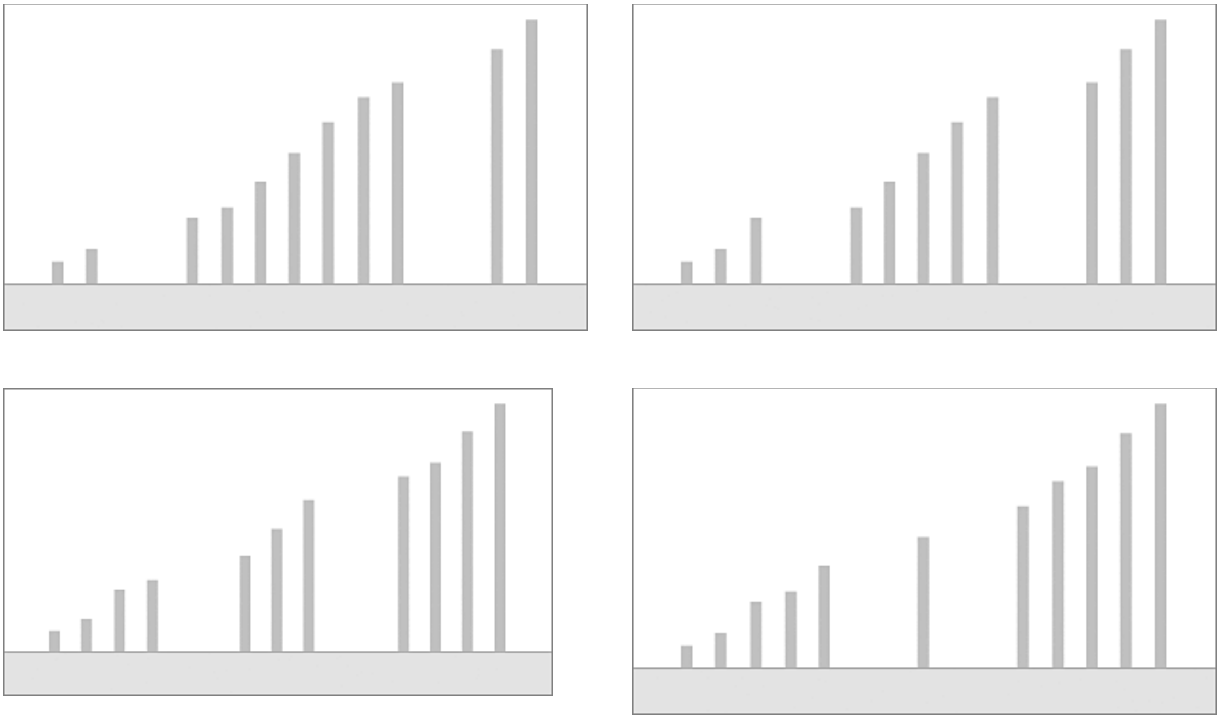
First, place your 11 noodles in order from shortest to longest on a new piece of paper or cardboard. Your arrangement should look something like this:



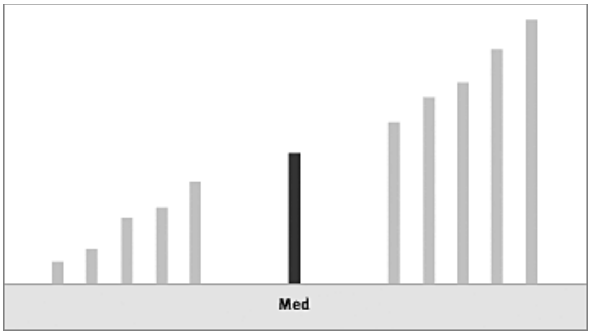
Next, remove two noodles at a time, one from each end, and put them to the side:



Part B, cont'd.



Continue this process until only one noodle remains. This noodle is the median. Label it "Med":



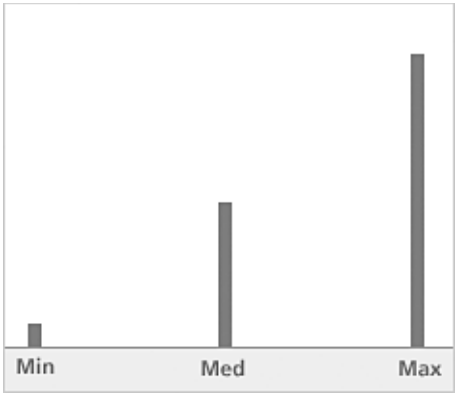
Problem B1. If you could see only the median noodle, what would you know about the other noodles? **[See Tip B1, page 121]**

Problem B2. If you could see only the median noodle, describe some information you would *not* know about the other noodles.

Part B, cont'd.

The Three-Noodle Summary

Now remove all the noodles except Min, Med, and Max.



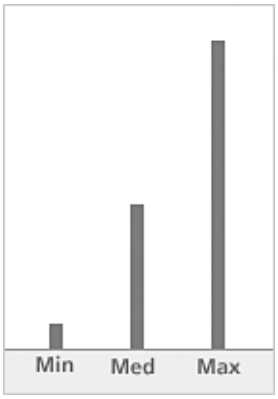
We'll call this display the "Three-Noodle Summary."

Problem B3. If you could see Min, Med, and Max, what would you know about the other noodles? Be specific about how this compares to Problem A3 (where you only knew Min and Max) and Problem B1 (where you only knew Med).

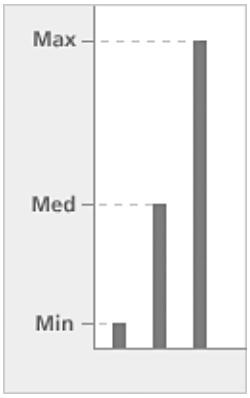
Problem B4. Describe some information you *still* wouldn't know about the other noodles from the Three-Noodle Summary.

The Three-Number Summary

Now let's convert the Three-Noodle Summary to the Three-Number Summary. If they're not already there, place the three noodles—Min, Med, and Max—in order on the horizontal axis, like this:



Next add a vertical number line, and mark the lengths of these three noodles:



Remove the noodles, and you're left with the Three-Number Summary:

Part B, cont'd.

Problem B5. If we call the length of the fourth noodle N4, how does N4 compare to Min, Med, and Max? What wouldn't you know about N4 if you only knew Min, Med, and Max?

Even Data Sets

In the previous example, it wasn't hard to find the median because there were 11 noodles—an odd number. For an odd number of noodles, the median is the noodle in the middle. But how do we find the median for an even number of noodles?

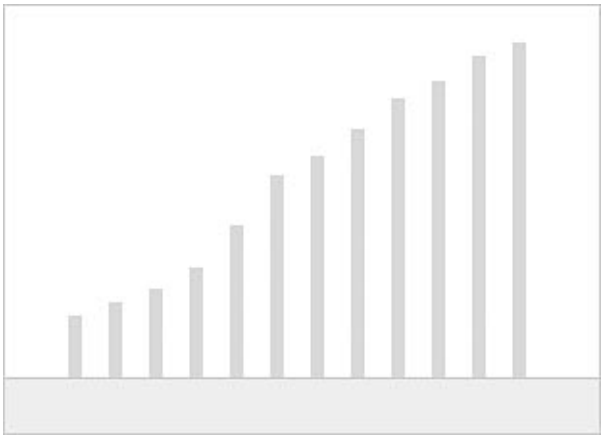
Add a 12th noodle, with a different length from the other 11 noodles, to the original collection. Arrange the noodles in order from shortest to longest.

Problem B6. Using the method of removing pairs of noodles (the longest and the shortest), try to determine the median noodle length. What happens?

For a non-interactive version of this activity, look at the following illustrations.

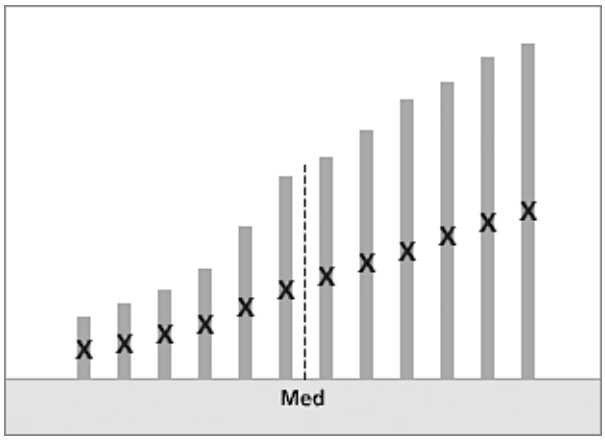
Try It Online! **www.learner.org**

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 4, Part B, Problem B6.



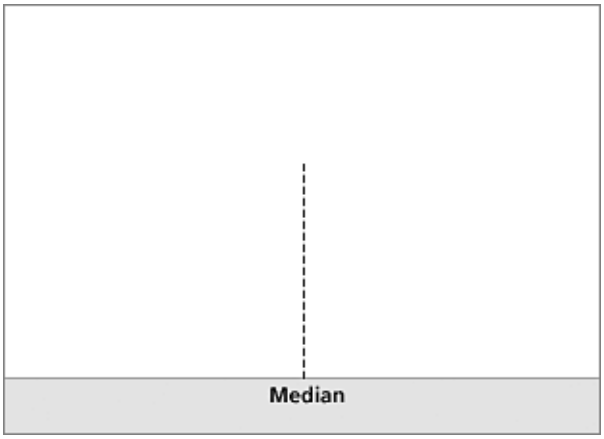
This time, there won't be one remaining noodle in the middle—there will be two! If you remove this middle pair, you'll have no noodles left.

Therefore, you'll need to draw a line midway between the two remaining noodles to play the role of the median. The length of this line should be halfway between the lengths of the two middle noodles:



Part B, cont'd.

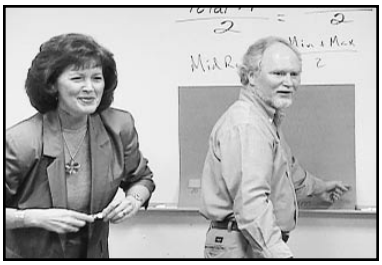
Problem B6, cont'd.



Move the middle pair aside, and you can see your new median:

Notice that this median still divides the set of noodles into two groups of the same size—the six noodles shorter than the median and the six noodles longer than the median. The major difference is that, this time, the median is not one of the original noodles; it was computed to divide the set into two equal parts.

Note: It is a common mistake to include this median in your data set when you've added it in this way. This median, however, is not part of your data set.



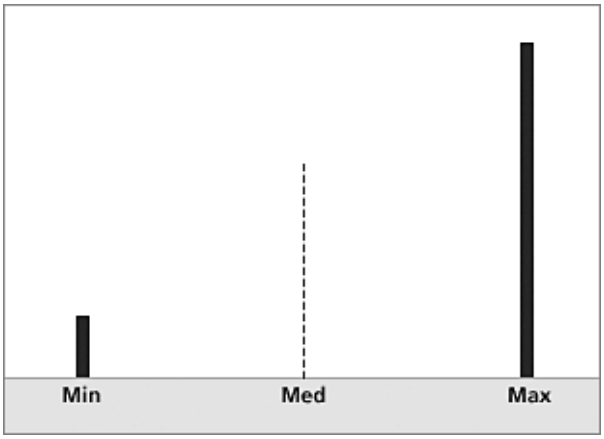
Video Segment (approximate time: 5:08-6:13): You can find this segment on the session video approximately 5 minutes and 8 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing.

In this video segment, participants discuss the process of finding the median of a data set with an even number of values (in this case $n = 20$). Watch this video segment to review the process you used in Problem B6 or if you would like further explanation.

Note: The data set used by the onscreen participants is different from the one provided above.

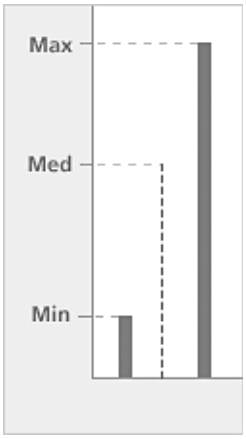
Problem B7. If you could see only the median of a set of 12, what would you know about the other noodles?

You can convert the Three-Noodle Summary for these 12 noodles to the Three-Number Summary in the same way you did it for the set of 11 noodles:

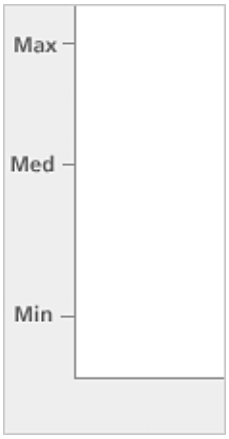


Part B, cont'd.

Add a vertical number line, and mark the lengths of the three noodles:



Remove the noodles, and you're left with the Three-Number Summary:



Review

As we have seen with the noodle examples, the median divides ordered numeric data into two groups, each with the same number of data values.

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

You can review the Three-Noodle Summary for odd and even data sets online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 4, Part B.

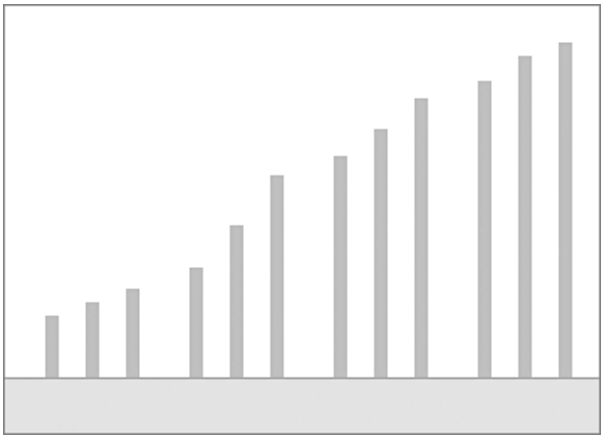
If you only know the Three-Number Summary (Min, Med, and Max) for a set of data, you can still glean quite a bit of information about the data. You know that all the data values are between Min and Max, and you know that Med divides the data into two groups of equal size. One group contains data values to the left of Med, and the other group contains data values to the right of Med. You also know that the group of values to the left of the median must be lower than (or equal to) the median in value, and that the group of values to the right of the median must be greater than (or equal to) the median in value.

Part C: Quartiles and the Five-Number Summary (35 min.)

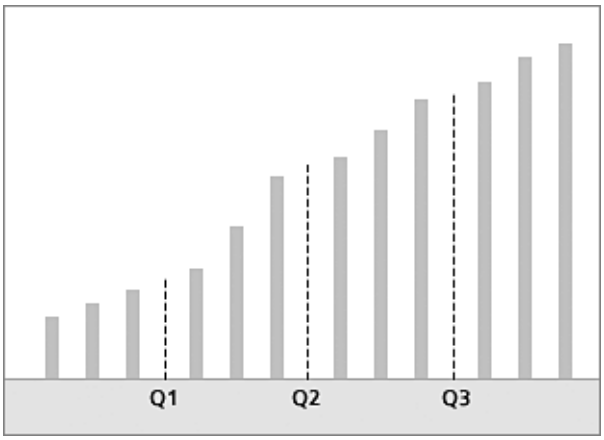
Quartiles

Let's go back to your 12 ordered noodles, arranged from shortest to longest on a new piece of paper or cardboard. As before, two of the noodles will be Min and Max. Now we're going to identify three noodles that divide the 12 (including Min and Max) into four groups of the same size.

First, divide your noodles into four groups with an equal number of noodles in each:



As with the Three-Noodle Summary for an even data set, we need to insert three extra lines, which we'll label Q1, Q2, and Q3, to divide and define the groups:



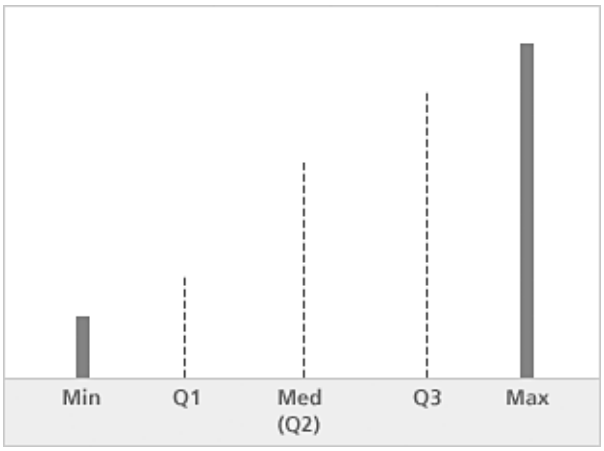
Note that Q2 is the median (Med) of this data set, since six noodles are to the left of Q2 and six are to the right.

Problem C1. What is the median of the six noodles to the left of Q2? What is the median of the six noodles to the right of Q2? **[See Tip C1, page 121]**

Q1, Q2, and Q3 are called quartiles, since they divide the noodles into four groups (i.e., quarters), with an equal number of noodles in each group. The line Q1 is the median of the six noodles to the left of Q2, and Q3 is the median of the six noodles to the right of Q2. Q2 is the median of the entire set of noodles.

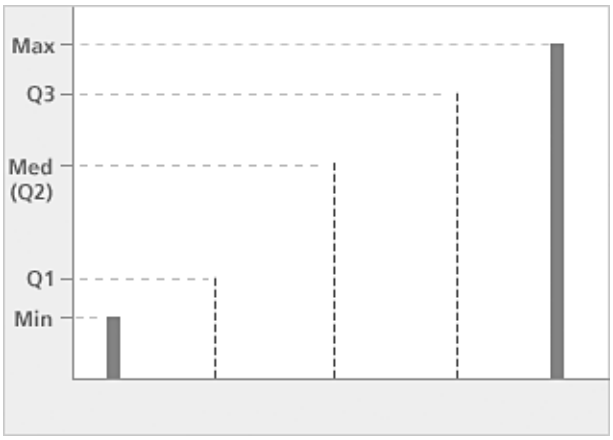
Part C, cont'd.

The Five-Noodle Summary consists of Min, Q1, Med (Q2), Q3, and Max:

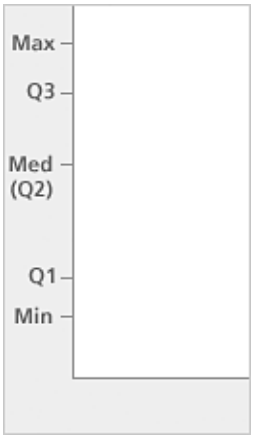


Problem C2. Using the information given in this Five-Noodle Summary, describe what you know about the 12 noodles. For example, what do you know about the ninth noodle, and what information are you still missing?

To convert the Five-Noodle Summary to the Five-Number Summary, use the same procedure you've followed throughout this session. Add a vertical number line so that you can indicate the lengths of the five noodles:



Remove the noodles, and you're left with the Five-Number Summary:



The number Q1 is called the *first* or *lower* quartile. The number Q3 is called the *third* or *upper* quartile.

Part C, cont'd.

Problem C3. If N4 is the length of the fourth noodle, what information would you know about N4 from the Five-Number Summary?

Problem C4. Ralph claims that the Five-Number Summary is enough to know that N4 is closer to Q1 than it is to Med. He says, "Since N4, N5, and N6 are all between Q1 and Med, N4 has to be closer to Q1 than it is to Med." Is his reasoning valid? Why or why not? **[See Tip C4, page 121]**

More Five-Number Summaries

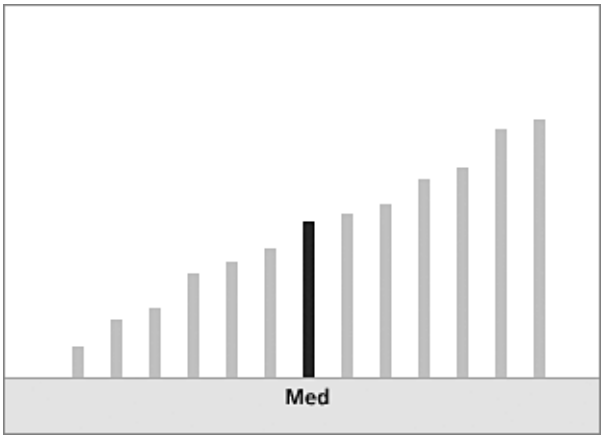
In the previous example, there were 12 noodles. Twelve is a convenient number of data values for introducing quartiles, because it is an even number and it is divisible by four. In this case, the quartiles separate the data into groups that each contain three values.

Quartiles always produce four groups of data with an equal number of data values in each group. But when the total number of data values is not divisible by four, it's trickier to determine exactly how many values will be in each of the four groups.

Determining quartiles is a two-step process:

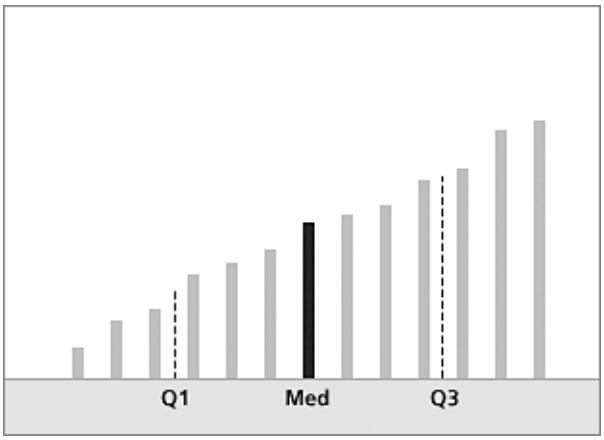
- First, find the median, or Med. Med divides the ordered data into two groups of equal size. One group contains data values to the left of Med, and the other contains data values to the right of Med.
- Next, find the median of the data values to the left of Med, which is the first quartile (Q1). Similarly, the third quartile (Q3) is the median of the data values to the right of Med.

Let's illustrate how this works for 13 data values (i.e., noodles). Since the total number of noodles is now odd, the median will be one of the original 13 noodles. Note that there is the same number of noodles to the left of Med as there is to the right of Med. Since you cannot divide the 13 noodles into two equal groups without splitting a noodle, take one noodle in the middle as the median and divide the other 12 noodles into two equal groups. This will occur whenever there is an odd number of noodles. The two equal groups will have exactly half of the noodles, with one noodle left in the middle as the median.



Part C, cont'd.

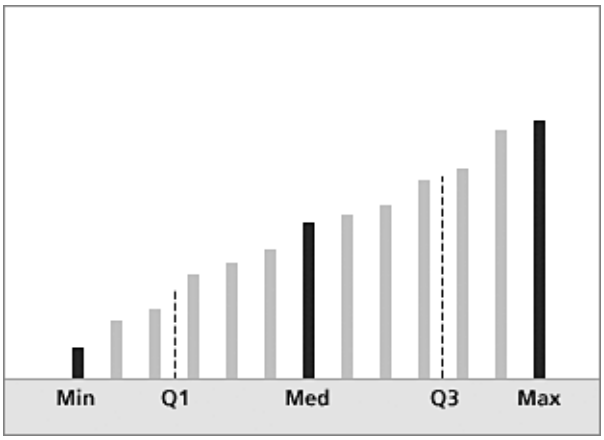
Now we find Q1, the median of the six noodles to the left of Med, and Q3, the median of the six noodles to the right of Med. Because there is an even number of noodles to the left and right of Med, Q1 and Q3 will be represented by lines between a pair of noodles.



Note that there are three noodles to the left of Q1, three noodles between Q1 and Med, three noodles between Med and Q3, and three noodles to the right of Q3. Also note that each group of three noodles is approximately one-fourth of the total of 13 noodles. As with the calculation of the median, the quartiles split each half of the noodles into two equal groups; if there is an odd number of noodles in a half, one will be left in the middle as the quartile.

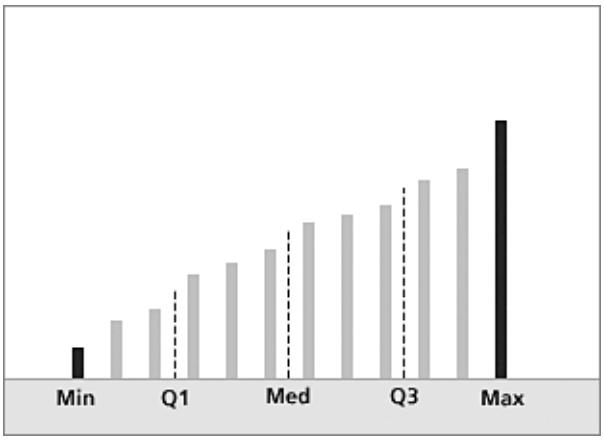
Review

Problem C5. Review how you would find the Five-Noodle Summary for a set of 13 noodles and a set of 12 noodles using the following illustrations:



Part C, cont'd.

Problem C5, 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 ww.learner.org/learningmath and find Session 4, Part C, Problem C5.

Problem C6. Explain how you would create a Five-Noodle Summary for 14 noodles. How many noodles are in each of the four groups? **[See Tip C6, page 121]**

Problem C7. Explain how you would create a Five-Number Summary for 15 noodles. How many numbers are in each of the four groups?

Take It Further

Problem C8. How many numbers are in each of the four groups if you started with 57 noodles? With 112 noodles? Can you find a rule that would allow you to determine the number of values in each group without creating a Five-Number Summary?

In general, the Five-Number Summary divides ordered numeric data into four groups**,** with each group having the same number of data values. If you know only the Five-Number Summary (Min, Q1, Med, Q3, and Max), these five values still give you a lot of information:

- All the data values are between Min and Max.
- Med divides the ordered data into two groups, with an equal number of values (approximately half) in each group:
  - One group contains data values to the left of Med.
  - One group contains data values to the right of Med.
- The quartiles divide the ordered data into four groups, with an equal number of values (approximately one-fourth) in each group:
  - One group contains values to the left of Q1 (and includes Min).
  - One group contains values between Q1 and Med.
  - One group contains values between Med and Q3.
  - One group contains values to the right of Q3 (and includes Max).

Problem C9. What information is learned from the interquartile range, the length of the interval between Q1 and Q3? Think about why this might be useful in describing the variation in your data.

Part D: The Box Plot (25 min.)

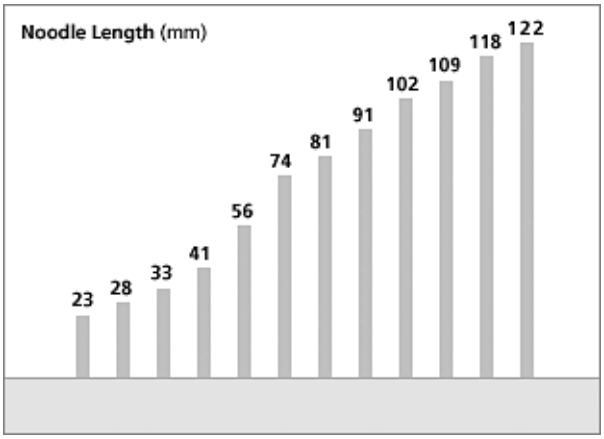
Five-Number Summary With Measurement Data

Now we'll look at how you can represent the Five-Number Summary graphically, using a box plot. For this activity, we will work with a set of 12 noodles with the following measurements (in millimeters):

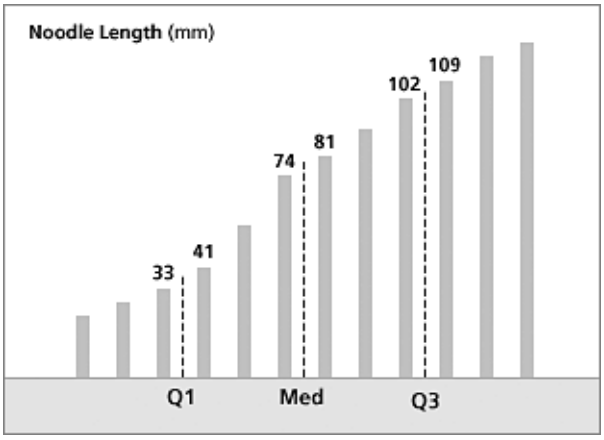
23 28 33 41 56 74 81 91 102 109 118 122

Problem D1. Why is it necessary to order the data before creating a Five-Number Summary?

Let's create a Five-Number Summary for this set of ordered data:



Determine Q1, Med, and Q3:



Part D, cont'd.

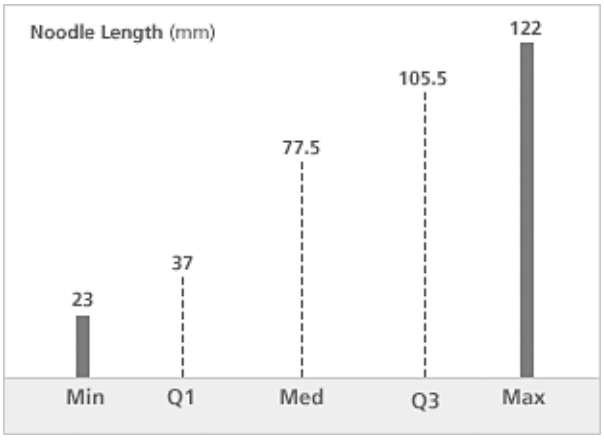
The lines representing Q1, Med, and Q3 each have lengths that are halfway between their adjacent noodles:

Q1 = (33 + 41) / 2 = 37

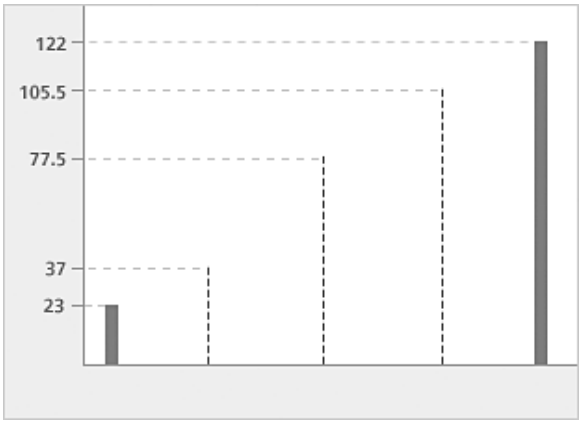
Med = (74 + 81) / 2 = 77.5

Q3 = (102 + 109) / 2 = 105.5

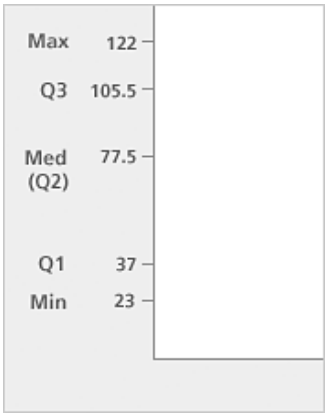
Here is the Five-Noodle Summary:



Add a vertical number line:



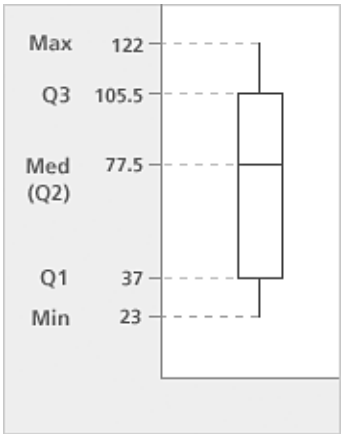
Here is the Five-Number Summary:



Part D, cont'd.

Drawing a Box Plot

Once we have the Five-Number Summary, we can display it using a kind of graph known as a box plot. Here is the box plot for the noodle data we've been using:



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

You can compare the noodle data as represented by the Five-Noodle Summary, the Five-Number Summary, and the box plot online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 4, Part D.

The box plot is also called a box-and-whiskers plot. Though it looks very different from previous graphs, it's just another way to represent the distribution of the data we've been working with all along:

- The lower whisker extends from Min to Q1. The length of this whisker indicates the range of the lowest (or, in this case, the shortest) fourth of the ordered data.
- The upper whisker extends from Q3 to Max. The length of this whisker indicates the range of the highest (or, in this case, the longest) fourth of the ordered data.
- The box (the rectangular portion of the graph) extends from Q1 to Q3, with a horizontal line segment indicating Med.
- The portion of the rectangle between Q1 and Med indicates the range of the second fourth of the ordered data.
- The portion of the rectangle between Med and Q3 indicates the range of the third fourth of the ordered data.
- The entire rectangle indicates the range of the middle half (the interquartile range) of the ordered data.

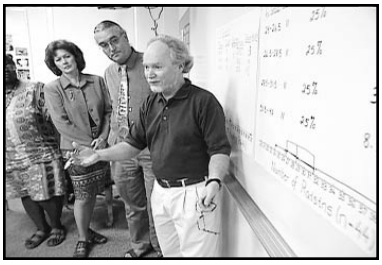
Note that the box plots can be drawn vertically or horizontally, depending on whether you display the Five-Number Summary along a vertical or a horizontal axis. **[See Note 2]**

Note 2. The Five-Number Summary uses intervals to describe the variation in different segments of your data. The longer the interval, the greater the variation. Some people will misinterpret a box plot. For example, given a box plot with the Q3-Max whisker considerably longer than the Min-Q1 whisker, one could think, "Wow, there are a lot more data in the highest interval than there are in the lowest interval." We're used to associating length with "how many" rather than "how far apart," and we forget that the same number of values falls within each of these intervals.

It is also important to note the difference between a histogram and a box plot, another potential source of confusion. To construct a histogram, you prescribe intervals of uniform length and then count how many data values fall within each interval. To determine the five numbers for the box plot, you do the reverse: prescribe how many data values you want in each interval and then determine the intervals.

Fathom Software, used by the onscreen participants, is helpful in creating graphical representations of data. You can use Fathom Software to complete Problems D2-D3. For more information, go to the Key Curriculum Press Web site at http://www.keypress.com/fathom/.

Part D, cont'd.



Video Segment (approximate time: 15:10-16:28): You can find this segment on the session video approximately 15 minutes and 10 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing.

In this video segment, Professor Kader introduces the process of building a box plot. Watch this segment to review the process or to help you draw the box plots for the following problem.

Note: The data set used by the onscreen participants is different from the one provided above.

Problem D2. Using the same scale for each plot, create a box plot for each of the data sets below, which we first saw in Session 2. Each is an ordered list of the number of raisins in a group of boxes from a particular brand. You may want to save your data for use in Session 6.

First create a box plot for the following raisin counts for boxes of Brand A raisins:

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

Then create a box plot for these raisin counts for boxes of Brand B raisins:

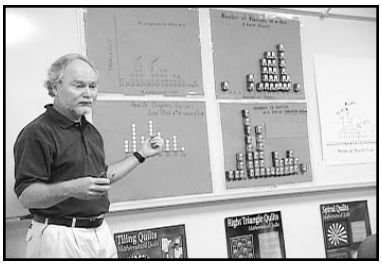
17 22 24 24 25 25 25 25 26 26 26

26 26 26 27 27 27 27 28 29 29 29

29 29 29 30 30

[See Tip D2, page 121]

Problem D3. Compare the two box plots from Problem D2 side by side. What conclusions can you draw about Brand A raisins in comparison to Brand B raisins, using only the box plots?



Video Segment (approximate time: 18:18-20:09): You can find this segment on the session video approximately 18 minutes and 18 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing.

In this video segment, Professor Kader and participants use the box plot to compare different brands of raisins. They then discuss the usefulness of the box plot as a summary of data. Watch this segment after completing Problem D3.

Note: The data set used by the onscreen participants is different from the one provided above.

Is the box plot more useful for making comparisons between different distributions than a line plot? Why or why not?

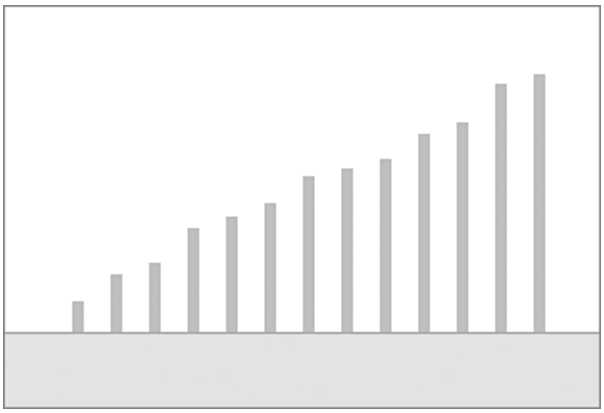
Part E: Finding the Five-Number Summary Numerically (30 min.)

Locating the Median From Ordered Data

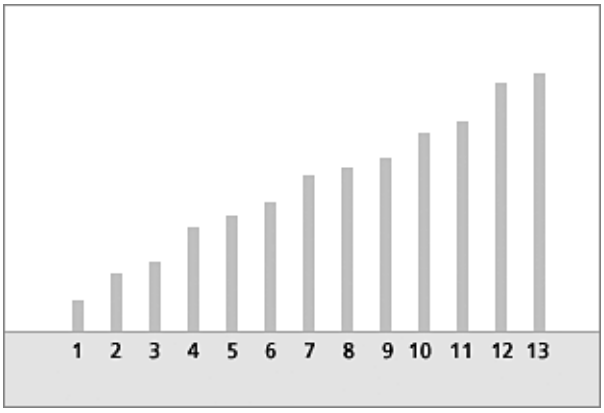
In Part D, we used noodles to help us visualize the concept of quartiles. In practice, however, the task of determining quartiles is treated strictly as a numerical problem. It is based on an ordered list of numerical measurements and the position of each measurement in the list. In Part E, we'll transition to this numerical approach. **[See Note 3]**

Remember the procedure for determining quartiles described earlier: First find the median; then find the first and third quartile values.

Let's begin with 13 noodles, arranged in ascending order:



Each noodle has a position in this ordered list: (1) indicates the shortest noodle, (2) the next shortest, and so on. The longest noodle is (13):

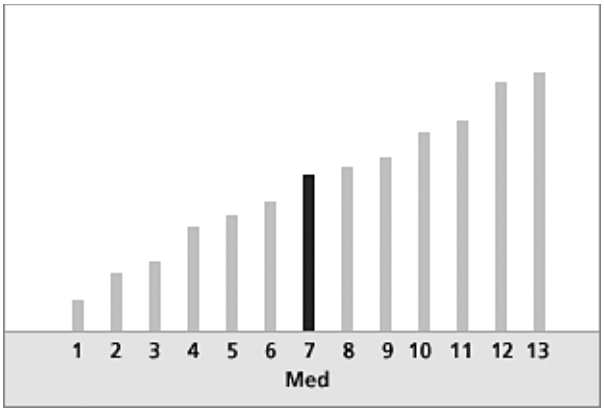


Note 3. You should be aware that several different algorithms are commonly used to determine the values of quartiles. The concept is the same, but the details of each method may differ.

For instance, the method we use in this course depends on a specific definition of "upper half" and "lower half." If you have an odd number of data values, you do not include the median in either half. This has become the popular method for teaching statistics in schools. It is also the method used in NCTM literature. Some statistics books or teaching materials, however, may use a slightly different method. For example, if you have an odd number of data values, you might include the median in both "halves." These two methods will sometimes produce the same or similar values for quartiles, but sometimes these values will be quite different, depending on the patterns and variation in your data.

Part E, cont'd.

The letter *n* is often used in statistics to indicate the number of data values in a set. In this case, there are *n* = 13 noodles, and 13 positions are indicated on the line above. The median is in position (7), because there are just as many positions (six) to the left of the median as there are to the right of the median:



The position of the median in an ordered list with *n* = 13 is (7). If there had been 14 items in the list, the position would have been halfway between positions (7) and (8), or (7.5). So if *n* = 14, the position of the median is (7.5).

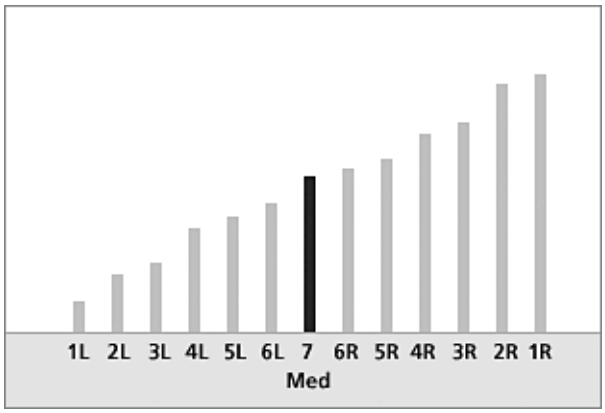
Problem E1. Find the position of the median for at least three other values of *n*. Then use this information to come up with a general mathematical rule for determining the position of the median if you know the number of items in an ordered list. **[See Tip E1, page 121]**

Calculating the Position of the Median

The general rule for determining the position of the median is that the median will always be in position (*n* + 1) / 2 in an ordered list. The positions can be indicated from smallest to largest (ascending order) or from largest to smallest (descending order). The median is in the same position and is the same value, regardless of the ordering method you use.

It is important to remember that this rule indicates the *position* of the median, and not the *value* of the median. The value of the median is the value at that position. In our 13-noodle example, *n* = 13, and the position of the median is determined by (13 + 1) / 2 = 14 / 2 = 7. So the median is in position (7), and the value of Med is the length of the seventh noodle.

Note that there are six noodles to the left (1L to 6L) and six noodles to the right (6R to 1R) of the median. To find the positions of the remaining quantities for the Five-Number Summary, it's convenient to label the noodles to the left of the median in ascending order and the noodles to the right of the median in descending order:



Again, notice that Med is in position (7) from each end of the ordered list, and notice that Min (the shortest noodle) and Max (the longest noodle) are each in position (1) on their respective ends of the ordered data.

Part E, cont'd.

Problem E2. What is the position of Q1 in this ordered list? **[See Tip E2, page 121]**

Problem E3. What is the position of Q3 in this ordered list?

Problem E4. Here is the ordered list of the numerical values for the 13 noodles and the corresponding position of each measurement from its respective end of the data:

13	23	28	33	41	56	74	81	91	102	118	122	127
(1L)	(2L)	(3L)	(4L)	(5L)	(6L)	(7)	(6R)	(5R)	(4R)	(3R)	(2R)	(1R)

Use the information from the position of the data in this ordered list and the results from Problems E2 and E3 to build the Five-Number Summary for this data. **[See Tip E4, page 121]**

Problem E5. Use the techniques you've learned in Part E to build the Five-Number Summary for the following set of measurements, where $n = 15$:

10	12	15	18	22	24	25	26	34	39	45	51
62	75	89									

Problem E6. Build the Five-Number Summary for the following set of measurements, where $n = 20$:

1	1	1	3	4	4	5	10	10	17	20	24
26	33	34	38	39	50	53	53				

[See Tip E6, page 121]

Problem E7. Here are the lengths of 20 pine needles, to the nearest millimeter, from Session 1, Problem H1:

117, 56, 48, 69, 71, 120, 111, 49, 68, 110, 109, 64, 93, 43, 109, 37, 93, 40, 86, 47

a. Determine the Five-Number Summary for these 20 measurements.

b. Draw a box plot for these 20 measurements.

c. Give a brief interpretation of this summary. What does it tell you about the lengths of the pine needles?

[See Tip E7, page 121]

Homework

Problem H1. Determine the Five-Number Summary for each of the remaining data sets of raisin counts from Session 2, and construct a box plot for each on the same scale as the ones you built in Problem D2. Then interpret the quantities in each Five-Number Summary. In other words, use your results to answer the question "How many raisins are in a half-ounce box of raisins?" for each brand.

Homework, cont'd.

Problem H1, cont'd.

Here are the raisin counts for boxes of Brand C raisins:

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

Here are the raisin counts for boxes of Brand D raisins:

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 the interpretations you made in Problems D2 and H1, which brand of raisins would you buy? Explain.

Problem H3. Consider the following data on sex, height, and arm span for 24 people from Session 1, Problem B3:

Sex	Height	Arm Span
Male	185	173
Female	160	161
Male	173	177
Female	170	170
Female	188	188
Male	184	196
Female	162	156
Female	170	162
Male	176	177
Female	166	165
Male	193	194
Male	178	178
Male	180	184
Female	162	159
Male	187	188
Male	186	200
Male	182	188
Female	160	157
Male	181	188
Male	192	188
Female	167	170
Female	176	173
Female	155	160
Female	162	161

Homework, cont'd.

Problem H3, cont'd.

a. Determine the Five-Number Summary and box plot for the 24 heights.

b. Determine the Five-Number Summary and box plot for the 24 arm spans.

c. Determine the Five-Number Summaries and box plots for the 12 males' and 12 females' heights. How do the box plots help you compare these two sets?

d. Determine the Five-Number Summaries and box plots for the males' and females' arm spans.

Take It Further

Problem H4. Describe how you would create a Four-Number Summary that divides the data into three groups with approximately one-third of the data in each group. Include instructions for determining the positions of T1 and T2 (the locations of the dividing points of the first and second thirds of the data).

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 and O'Connor, William (March, 1999). "Sticks to the Roof of Your Mouth?," *Mathematics Teaching in the Middle School, 4* (6), 404-411.

Kader, Gary and Perry, Mike (Summer, 1996). "To Boxplot or Not To Boxplot?," *Teaching Statistics, 18* (2), 39-41.

Tips

Part B: The Median and the Three-Number Summary

Tip B1. What would knowing the median tell you about each of the first five (the shortest five) noodles? What would it tell you about each of the last five (the longest five) noodles?

Part C: Quartiles and the Five-Number Summary

Tip C1. The median divides the set equally, so the median in a set of six noodles is the value that has three noodles to the left of it and three noodles to the right.

Tip C4. Try to build a data set that shows whether or not Ralph's claim is valid.

Tip C6. Remember that the median of a group may be represented by a noodle or by a line drawn halfway between two noodles. Since a quartile is the median of half the data, it may also be represented by a noodle or by a line drawn halfway between two noodles.

Part D: The Box Plot

Tip D2. Start by listing the position for each value in the data set. For example, in the set of Brand A raisins, the value 23 is in the first position, 25 is in the second position, the second 25 is in the third position, and so forth.

Part E: Finding the Five-Number Summary Numerically

Tip E1. Try consecutive numbers like 10, 11, and 12. To get you started, if $n = 10$, the median will be halfway between the fifth and sixth items, so the position of the median is (5.5).

Tip E2. Remember that you should only consider noodles to the left of the median. Do not include the median itself in this count.

Tip E4. Remember that a summary value that lies in a position halfway between two items in an ordered list is the average of the adjacent pair of values.

Tip E6. Ignore the *values* of the data when finding the positions of the median and quartiles. It is possible for the values surrounding the median and quartiles to be identical.

Tip E7. Don't forget that in order to build a Five-Number Summary or a box plot, you will need to order the list first!

Solutions

Part A: Min, Max, and the Two-Number Summary

Problem A1. You would know that the lengths of the other nine noodles must be between the lengths of these two; in other words, none of the other nine noodles can be shorter than Min, and none of them can be longer than Max.

Problem A2. The length of noodle N4 must be between Min and Max.

Problem A3. You would not know the mean length or the median length. You would not know whether the remaining nine noodles were closer to Min or to Max—only that they were between those values.

Problem A4. If you had the actual noodles or knew their lengths, you could use the mean as a "typical" value, which you find by adding the lengths of all 11 noodles and dividing the sum by 11. You could also use the median—the noodle in the center of the ordered list (i.e., the sixth noodle). However, if you only had the information from the Two-Number Summary, your best answer would be the average of Max and Min. This number, which is sometimes called the midrange, can turn out to be very far away from the mean and median, depending on the distribution of the noodles.

Part B: The Median and the Three-Number Summary

Problem B1. You would know that there must be exactly five noodles shorter than the median noodle and five noodles longer than the median noodle.

Problem B2. You would not know the actual values of any of the other noodles: The five shorter noodles could be extremely short, the five longer noodles could be many feet long, they could all be fairly close in size to the median, etc. You would also not know or be able to estimate the maximum or minimum length of the other noodles.

Problem B3. You would know that all of the noodles are between Min and Max, and you can divide the noodles into two equal groups: five that are shorter than Med (including Min) and five that are longer than Med (including Max). This information gives you two specific intervals that contain an equal number of noodles, and all of the noodles are contained in these intervals. This is different from Problem A3, where you knew nothing about the size of the noodles between Min and Max, and from Problem B1, where you knew nothing about the upper and lower boundaries of your data set.

Problem B4. You still wouldn't know the lengths of the noodles in the two intervals between Min and Med or between Med and Max. These noodles could be very close to Med, very close to the extreme values, evenly spread within the intervals, or something else entirely. There is no way to know without more information.

Problem B5. You would know that N4 must be larger than Min, smaller than Med, and smaller than Max. This is true because N6 is the median, and N4 must be smaller than N6. You still wouldn't know N4's actual value or whether N4 was closer to Min or to Med. (A common mistake is to claim that N4 must be closer to Med than it is to Min. This is not necessarily true, since the values of N2 through N5 can be anywhere in the interval between Min and Med; for example, they could all be very close to Min.)

Problem B6. There are two noodles left, the sixth and the seventh. Neither of these two noodles can serve as the median, so we need to do something else.

Problem B7. You would know that there are six noodles that are shorter than the median and six that are longer.

Solutions, cont'd.

Part C: Quartiles and the Five-Number Summary

Problem C1. For six noodles, the median is located between the third and fourth noodle. For the six noodles to the left of Q2, this median will be Q1. Similarly, the median of the six noodles to the right of Q2 will be Q3.

Problem C2. You would know that all of the lengths are between Min and Max, and that Med (Q2) divides the ordered data into two equal-sized groups. Six noodles will be shorter than Med, and six will be longer. The quartiles then divide the ordered data into four equal-sized groups. The first group contains three noodles shorter than Q1; these three noodles must have lengths the size of or larger than Min and smaller than Q1. The second group contains three noodles that are longer than Q1 but shorter than Med. The third group contains three noodles that are longer than Med but shorter than Q3. The final group contains three noodles that are longer than Q3 and the size of or smaller than Max. (For example, the ninth noodle is longer than Med and shorter than Q3.) You still don't know how the three noodles in each group are distributed—only the ends of each interval. (For example, you don't know whether the ninth noodle is closer to Q3 or to Med.)

Problem C3. You would know that N4 is larger than Q1, the first quartile, and that it is shorter than Med, the median.

Problem C4. No, Ralph's reasoning is not necessarily valid. Here is a sample data set of noodle lengths, measured to the nearest millimeter: 30, 35, 38, 60, 61, 62, 64, 67, 70, 75, 90, 96. The fourth noodle, N4, has a length of 60 mm. The first quartile, Q1, is (38 + 60) / 2 = 49 mm. The median is (62 + 64) / 2 = 63 mm. In this set, N4 is closer to Med than to Q1. Remind Ralph that the information in the Five-Number Summary, while valuable, does not tell us anything about the actual *values* in each interval. Ralph's claim would only be valid if the data are equally spaced, for example if each length was a multiple of 10.

Problem C5. Review your answer online with the Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 4, Part C, Problem C5.

Problem C6. First, find the median, between the seventh and eighth noodles. The first quartile is the median of the seven shortest noodles, which is the fourth noodle. The third quartile is the median of the seven longest noodles, which is the 11th noodle. There will be three noodles in each of the four groups.

Problem C7. First, find the median, which will be the eighth noodle. The first quartile is the median of the seven shortest noodles, which is the fourth noodle. The third quartile is the median of the seven longest noodles, which is the 12th noodle. There will be three noodles in each of the four groups.

Problem C8. If you started with 57 noodles, there would be 14 noodles in each group (the median is the 29th noodle). If you started with 112 noodles, there would be 28 noodles in each group (the median is between the 56th and 57th noodles). One possible rule is to take the number of noodles, divide by four, and then round down if you have a fractional result.

Problem C9. The interquartile range contains the center 50% of the data. This is a useful interval for describing variation; if the interquartile range is small compared to the overall range (from Min to Max), it suggests that there are a lot of extreme values in the data. If the interquartile range is wide compared to the overall range, it suggests that there are few extreme values and that the data are pretty tightly grouped.

Solutions, cont'd.

Part D: The Box Plot

Problem D1. Since the median and quartiles require separating the data into halves that are larger or smaller than a central value, it is necessary to order the data. If the data are unordered, it is much more difficult to find the value that splits the list into two equal groups.

Problem D2. To create a box plot, first create a Five-Number Summary for each data set:

a. For Brand A, here is the Five-Number Summary:

Min = 23

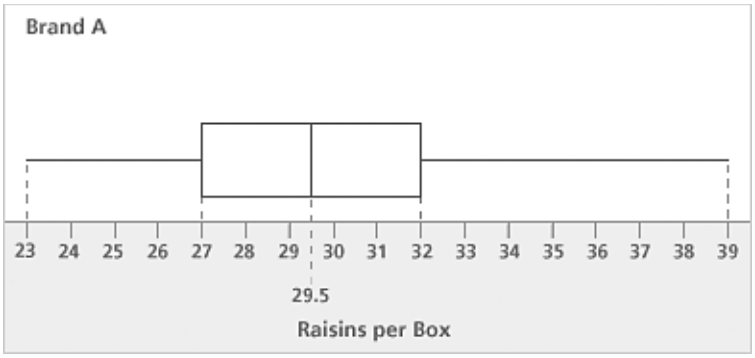
Q1 = 27

Med = 29.5

Q3 = 32

Max = 39

Here is the box plot:



b. For Brand B, here is the Five-Number Summary:

Min = 17

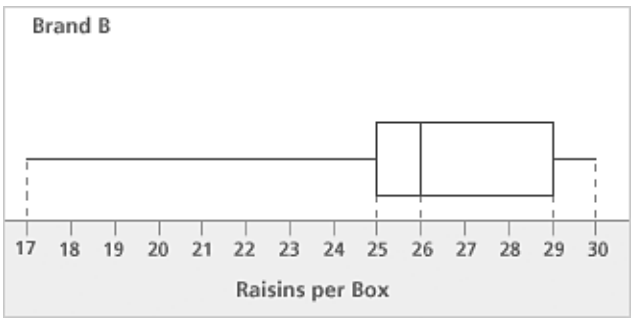
Q1 = 25

Med = 26

Q3 = 29

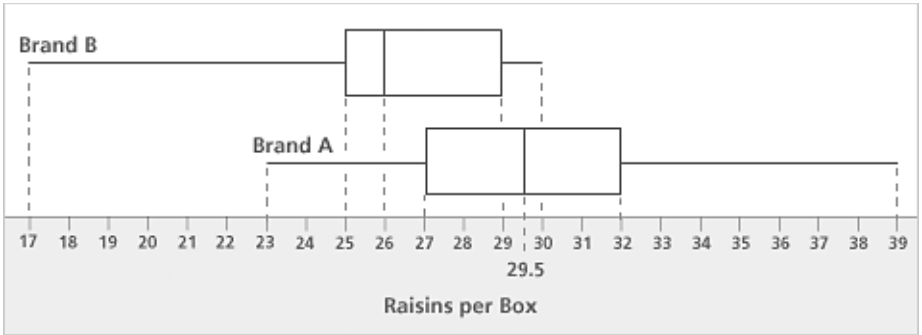
Max = 30

Here is the box plot:



Solutions, cont'd.

Problem D3. Placing the box plots side by side clearly shows that a large number of Brand A boxes have more raisins than Brand B boxes. The interquartile range is a little wider for Brand A, and the top 25% of Brand A boxes are all higher than Brand B's maximum. This suggests strongly that Brand A, on average, has more raisins in a typical box than Brand B.



Part E: Finding the Five-Number Summary Numerically

Problem E1. Here are some examples. If there were 15 raisins (n = 15), the median would be in position (8). If n = 16, the median would be in position (8.5). If n = 17, the median would be in position (9). A general mathematical rule is that the position of the median is (n + 1) / 2, where n is the number of items in the list.

Problem E2. Since six is an even number, this is a case where you would need to draw a line to represent the position of Q1, the median of the six noodles to the left of Med. Using the formula (n + 1) / 2 from Problem E1 gives us (6 + 1) / 2 = 7 / 2 = 3.5. Therefore, position (3.5L), halfway between positions (3L) and (4L), is the position (though not the value) of Q1.

Problem E3. Again, you'll need to draw a line to represent the position of Q3. As in Problem E2, the formula (n + 1) / 2 gives us (6 + 1) / 2 = 7 / 2 = 3.5. Therefore, position (3.5R), halfway between positions (3R) and (4R), is the position (though not the value) of Q3.

Problem E4. Here is the Five-Number Summary:

- Min is in position (1L); Min = 13.
- Max is in position (1R); Max = 127.
- Med is in position (13 + 1) / 2 = (7); Med = 74.
- There are six positions to the left of (7), so Q1 is in position (6 + 1) / 2 = (3.5L). The value of Q1 is (28 + 33) / 2; Q1 = 30.5.
- There are six positions to the right of (7), so Q3 is in position (6 + 1) / 2 = (3.5R). The value of Q3 is (102 + 118) / 2; Q3 = 110.

Solutions, cont'd.

Problem E5. First, number the positions as you did in Problem E4. The center position will be marked with an (8). Here is the Five-Number Summary:

- The minimum is in position (1L); Min = 10.
- The maximum is in position (1R); Max = 89.
- The median is in position (15 + 1) / 2 = (8); Med = 26.
- The first quartile is in position (7 + 1) / 2 = (4L); Q1 = 18.
- The third quartile is in position (4R); Q3 = 51.

Problem E6. Again, number the positions as you did in Problem E4. This time, there will be two values numbered (10) in the center of the ordered list. Here is the Five-Number Summary:

- The minimum is in position (1L); Min = 1.
- The maximum is in position (1R); Max = 53.
- The median is in position (20 + 1) / 2 = (10.5), which means it is the average of the two values numbered (10), or (17 + 20) / 2; Med = 18.5.
- The first quartile is in position (10 + 1) / 2 = (5.5L), which is (4 + 4) / 2; Q1 = 4.
- The third quartile is in position (5.5R), which is (34 + 38) / 2; Q3 = 36.

Problem E7. The ordered list is as follows:

37, 40, 43, 47, 48, 49, 56, 64, 68, 69, 71, 86, 93, 93, 109, 109, 110, 111, 117, 120.

Here is the Five-Number Summary:

a. Min = 37

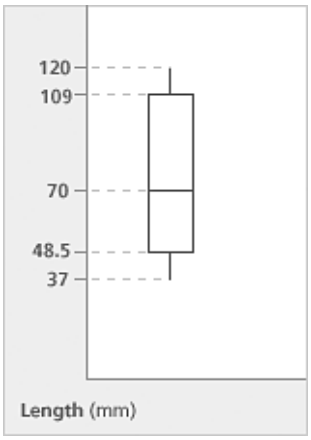
   Q1 = 48.5

   Med = 70

   Q3 = 109

   Max = 120

b. Here is the box plot:



Solutions, cont'd.

Problem E7, cont'd.

c. Based on these measurements:

- All pine needles have lengths between 37 mm and 120 mm.
- Approximately half the pine needles have lengths less than 70 mm.
- Approximately half the pine needles have lengths greater than 70 mm.
- Approximately half the pine needles have lengths between 48.5 mm and 109 mm.
- The widest range of needle lengths seems to be in the third quartile, where 25% of the needles are between 70 mm and 109 mm.
- The longest and shortest needles fall in very tight ranges; the longest 25% of needles are between 109 and 120 mm, and the shortest 25% are between 37 mm and 48.5 mm.

Homework

Problem H1. For Brand C, there are 28 data entries. Here is the Five-Number Summary:

Min = 25

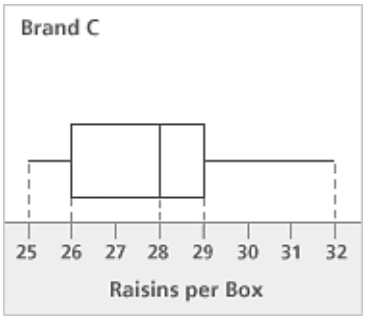
Q1 = 26

Med = 28

Q3 = 29

Max = 32

Here is the box plot:



The box plot shows that the center 50% of the data lies between 26 and 29, with 25% of the data falling between 28 and 29. Compared to other brands, Brand C has less variation, although there are a few boxes that have 30 or more raisins.

For Brand D, there are 36 data entries. Here is the Five-Number Summary:

Min = 23

Q1 = 27

Med = 29

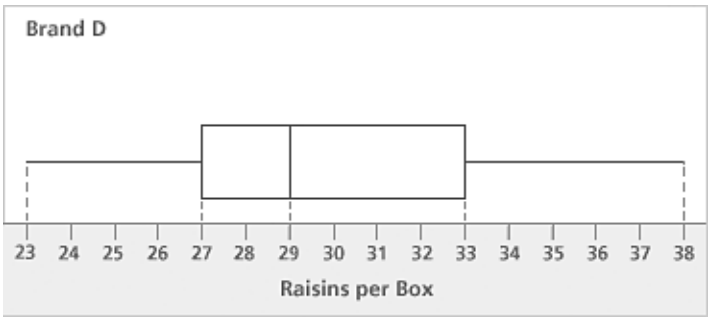
Q3 = 33

Max = 38

Solutions, cont'd.

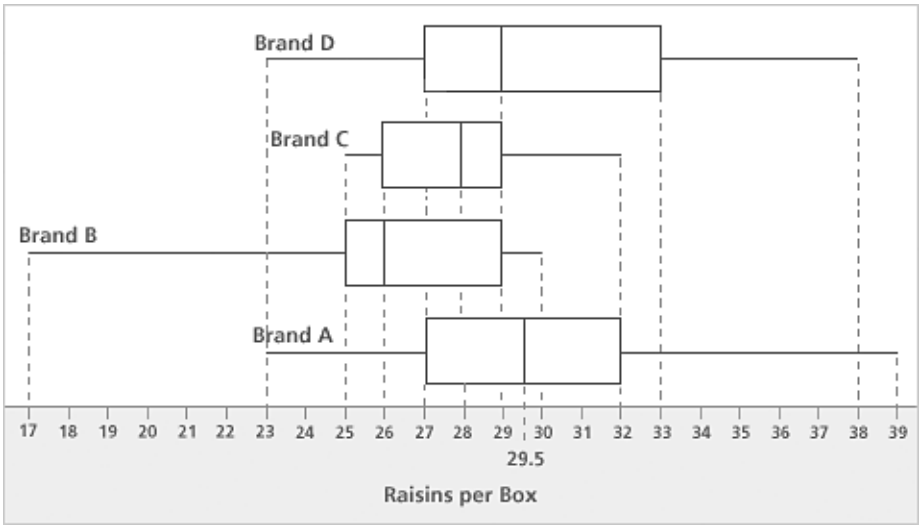
Problem H1, cont'd.

Here is the box plot:



The box plot shows that the center 50% of the data lies between 27 and 33, with 25% of the data falling between 29 and 33. Compared to other brands, Brand D has a large number of boxes with 30 or more raisins, but it also has far greater variation than Brand C, with boxes of as few as 23 raisins.

Problem H2. Answers will vary. The comparison of four box plots generally suggests that Brands A and D offer the greatest chance for a lot of raisins in a box, although Brand C offers the most consistency and the highest minimum number of raisins in a box. The answer to this question really depends on the goals of the individual purchasing the raisins!



Solutions, cont'd.

Problem H3. The first step is to order the lists from lowest to highest. Please note: Although each list is ordered from lowest to highest, the height and arm span measurements at the same position in the ordered list do not correspond to the same individual. (For example, the lowest height and lowest arm span are not necessarily from the same person.) Females are marked in **bold.**

Heights

155 160 160 162 162 162 166 167 170 170 173 **176** 176 178 180 181 182 184 185 186 187 **188** 192 193

Arm Spans

156 157 159 160 161 161 162 165 170 170 173 **173** 177 177 178 184 188 188 188 188 **188** 194 196 200

a. Here is the Five-Number Summary for the 24 heights:

Min = 155

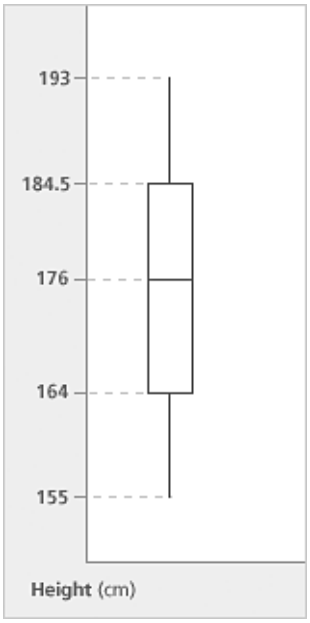
Q1 = 164

Med = 176

Q3 = 184.5

Max = 193

Here is the box plot:



b. Here is the Five-Number Summary for the 24 arm spans:

Min = 156

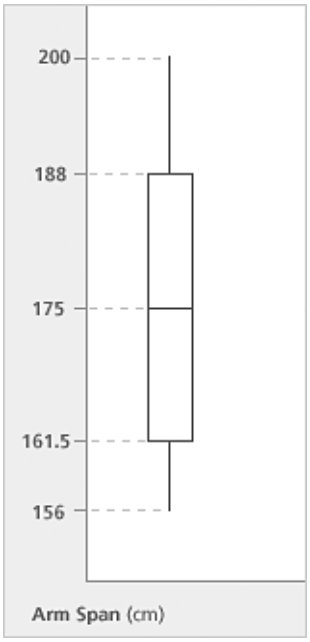
Q1 = 161.5

Med = 175

Q3 = 188

Max = 200

Here is the box plot:



Solutions, cont'd.

Problem H3, cont'd.

c. Here are the Five-Number Summaries:

Males' Heights

Min = 173

Q1 = 179

Med = 183

Q3 = 186.5

Max = 193

Females' Heights

Min = 155

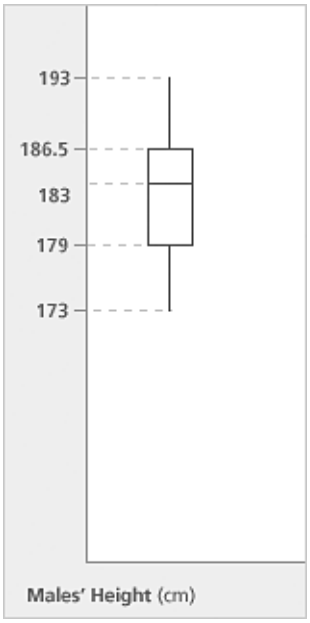
Q1 = 161

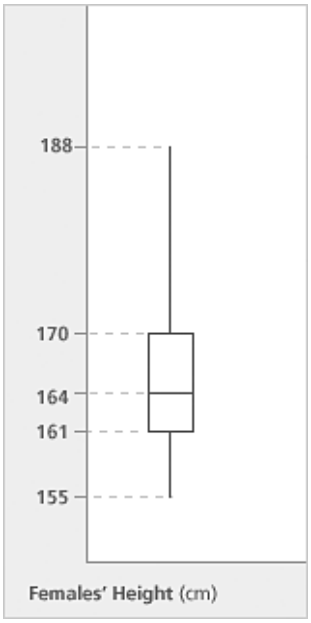
Med = 164

Q3 = 170

Max = 188

Here are the box plots:





Solutions, cont'd.

Problem H3, cont'd.

Comparing the box plots clearly shows that the males' heights are significantly greater than the females'. In particular, the third quartile value of females' heights was shorter than the minimum of males' heights, which shows that, in this survey, at least 75% of the females were shorter than the shortest male. However, the maximum height of a female is fairly close to the maximum height of a male, primarily because there was one very tall female!

d. Here are the Five-Number Summaries:

Males' Arm Spans

Min = 173

Q1 = 177.5

Med = 188

Q3 = 191

Max = 200

Females' Arm Spans

Min = 156

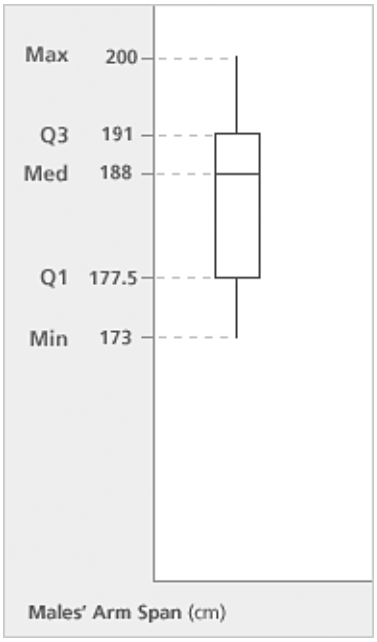
Q1 = 159.5

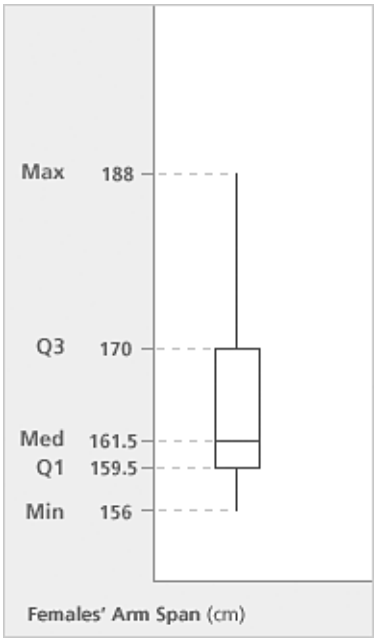
Med = 161.5

Q3 = 170

Max = 188

Here are the box plots:





Solutions, cont'd.

Problem H4. The Four-Number Summary comprises the maximum, the minimum, and two values T1 and T2, which mark the endpoints of the first and second thirds of your data. The locations of T1 and T2 are determined by *n*, the number of values in the data set, and are also based on whether *n* is a multiple of three, one more than a multiple of three, or two more than a multiple of three.

If *n* is a multiple of three, then the position of T1 is (n / 3 + 1/2) and the position of T2 is ($2n$ / 3 + 1/2). For example, if $n = 12$, T1 will be between the fourth and fifth data value (i.e., position [4.5]), and T2 will be between the eighth and ninth data value (i.e., position [8.5]). In this example, each of the three groups contains exactly four values.

If *n* is one more than a multiple of three, then the position of T1 is ($n + 2$) / 3, and the position of T2 is ($2n + 1$) / 3. For example, if $n = 13$, T1 will be the fifth position and T2 will be the ninth position. In this example, each group contains five values, if you include the endpoints.

If *n* is two more than a multiple of three, then the position of T1 is ($n + 1$) / 3, and the position of T2 is ($2n + 2$) / 3. For example, if $n = 14$, T1 will be the fifth position and T2 will be the 10th position. In this example, each group contains four values, if you don't include the endpoints.

Other answers are also possible, depending on whether you include the endpoints of the intervals.