Averages and Variation

3

Section 3.3

Percentiles and Box-and-Whisker Plots

Focus Points

- Interpret the meaning of percentile scores.
- Compute the median, quartiles, and five-number summary from raw data.
- Make a box-and-whisker plot. Interpret the results.
- Describe how a box-and-whisker plot indicates spread of data about the median.

Percentiles and Box-and-Whisker Plots

We've seen measures of central tendency and spread for a set of data. The arithmetic mean $\bar{x}$ and the standard deviation s will be very useful in later work.

However, because they each utilize every data value, they can be heavily influenced by one or two extreme data values.

In cases where our data distributions are heavily skewed or even bimodal, we often get a better summary of the distribution by utilizing relative position of data rather than exact values.

Percentiles and Box-and-Whisker Plots

We know that the median is an average computed by using relative position of the data.

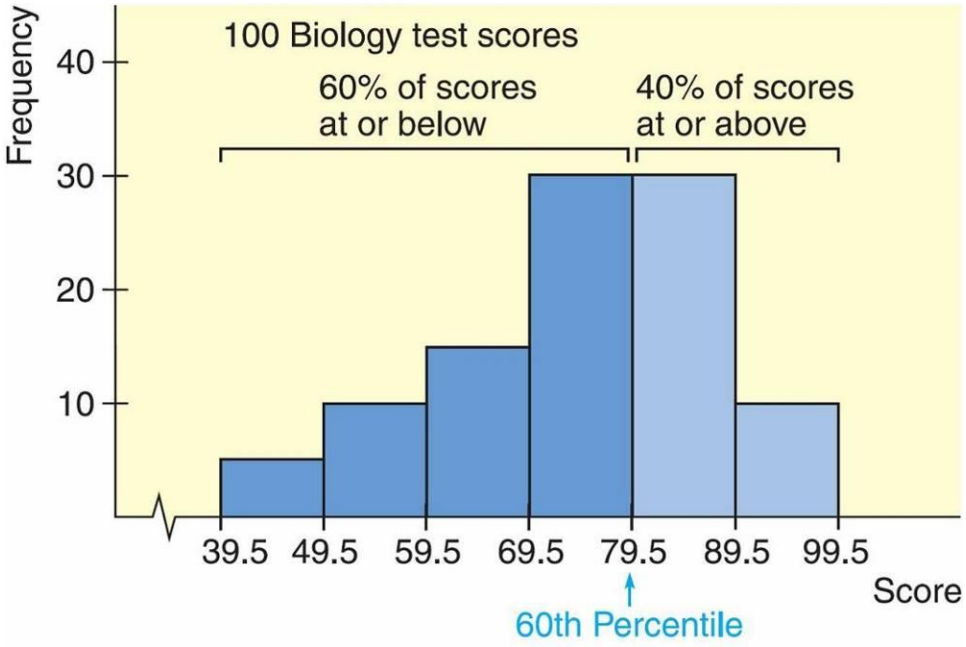
If we are told that 81 is the median score on a biology test, we know that after the data have been ordered, 50% of the data fall at or below the median value of 81.

The median is an example of a *percentile*; in fact, it is the 50th percentile. The general definition of the *P*th percentile follows.

For whole numbers P (where $1 \leq P \leq 99$), the Pth **percentile** of a distribution is a value such that $P\%$ of the data fall at or below it and $(100 - P)\%$ of the data fall at or above it.

Percentiles and Box-and-Whisker Plots

In Figure 3-3, we see the 60th percentile marked on a histogram. We see that 60% of the data lie below the mark and 40% lie above it.



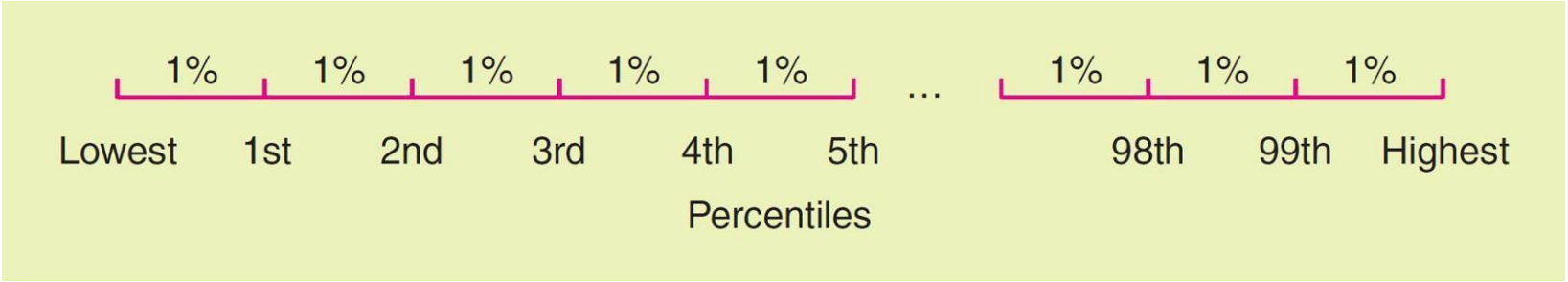
A Histogram with the 60th Percentile Shown

Figure 3-3

Percentiles and Box-and-Whisker Plots

There are 99 percentiles, and in an ideal situation, the 99 percentiles divide the data set into 100 equal parts. (See Figure 3-4.)

However, if the number of data elements is not exactly divisible by 100, the percentiles will not divide the data into equal parts.



Percentiles

Figure 3-4

Percentiles and Box-and-Whisker Plots

There are several widely used conventions for finding percentiles. They lead to slightly different values for different situations, but these values are close together.

For all conventions, the data are first *ranked* or ordered from smallest to largest. A natural way to find the *P*th percentile is to then find a value such that *P*% of the data fall at or below it.

This will not always be possible, so we take the nearest value satisfying the criterion. It is at this point that there is a variety of processes to determine the exact value of the percentile.

Percentiles and Box-and-Whisker Plots

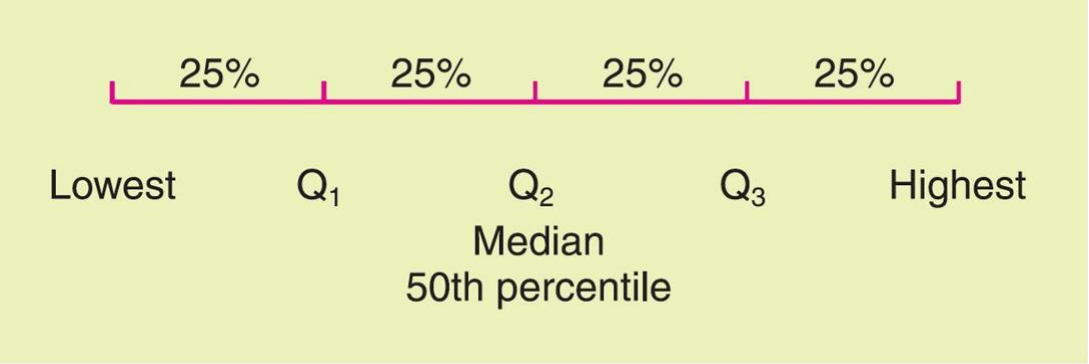
We will not be very concerned about exact procedures for evaluating percentiles in general.

However, *quartiles* are special percentiles used so frequently that we want to adopt a specific procedure for their computation.

Quartiles are those percentiles that divide the data into fourths.

Percentiles and Box-and-Whisker Plots

The *first quartile* Q_1 is the 25th percentile, the *second quartile* Q_2 is the median, and the *third quartile* Q_3 is the 75th percentile. (See Figure 3-5.)



Quartiles

Figure 3-5

Again, several conventions are used for computing quartiles, but the convention on next page utilizes the median and is widely adopted.

Percentiles and Box-and-Whisker Plots

Procedure:

HOW TO COMPUTE QUARTILES

1. Order the data from smallest to largest.
2. Find the median. This is the second quartile.
3. The first quartile Q_1 is then the median of the lower half of the data; that is, it is the median of the data falling *below* the Q_2 position (and not including Q_2).
4. The third quartile Q_3 is the median of the upper half of the data; that is, it is the median of the data falling *above* the Q_2 position (and not including Q_2).

Percentiles and Box-and-Whisker Plots

In short, all we do to find the quartiles is find three medians. The median, or second quartile, is a popular measure of the center utilizing relative position.

A useful measure of data spread utilizing relative position is the *interquartile range (IQR)*. It is simply the difference between the third and first quartiles.

Interquartile range = $Q_3 - Q_1$

The interquartile range tells us the spread of the middle half of the data. Now let's look at an example to see how to compute all of these quantities.

Example 9 – *Quartiles*

In a hurry? On the run? Hungry as well? How about an ice cream bar as a snack? Ice cream bars are popular among all age groups.

Consumer Reports did a study of ice cream bars. Twenty-seven bars with taste ratings of at least "fair" were listed, and cost per bar was included in the report.

Just how much does an ice cream bar cost? The data, expressed in dollars, appear in Table 3-4.

0.99	1.07	1.00	0.50	0.37	1.03	1.07	1.07
0.97	0.63	0.33	0.50	0.97	1.08	0.47	0.84
1.23	0.25	0.50	0.40	0.33	0.35	0.17	0.38
0.20	0.18	0.16					

Cost of Ice Cream Bars (in dollars)

Table 3-4

Example 9 – *Quartiles*

cont'd

As you can see, the cost varies quite a bit, partly because the bars are not of uniform size.

(a) Find the quartiles.

Solution:

We first order the data from smallest to largest. Table 3-5 shows the data in order.

0.16	0.17	0.18	0.20	0.25	0.33	0.33	0.35
0.37	0.38	0.40	0.47	0.50	0.50	0.50	0.63
0.84	0.97	0.97	0.99	1.00	1.03	1.07	1.07
1.07	1.08	1.23					

Ordered Cost of Ice Cream Bars (in dollars)

Table 3-5

Example 9 – *Solution*

cont'd

Next, we find the median.

Since the number of data values is 27, there are an odd number of data, and the median is simply the center, or 14^{th}, value.

The value is shown boxed in Table 3-5.

$$\text{Median} = Q_2 = 0.50$$

There are 13 values below the median position, and Q_1 is the median of these values.

Example 9 – *Solution*

cont'd

It is the middle, or seventh, value and is shaded in Table 3-5.

$$\text{First quartile} = Q_1 = 0.33$$

There are also 13 values above the median position. The median of these is the seventh value from the right end.

This value is also shaded in Table 3-5.

$$\text{Third quartile} = Q_3 = 1.00$$

Example 9 – *Quartiles*

cont'd

(b) Find the interquartile range.

Solution:

$$IQR = Q_3 - Q_1$$

$$= 1.00 - 0.33$$

$$= 0.67$$

This means that the middle half of the data has a cost spread of 67¢.

Box-and-Whisker Plots

Box-and-Whisker Plots

The quartiles together with the low and high data values give us a very useful *five-number summary* of the data and their spread.

Five-Number Summary

Lowest value, Q_1, median, Q_3, highest value

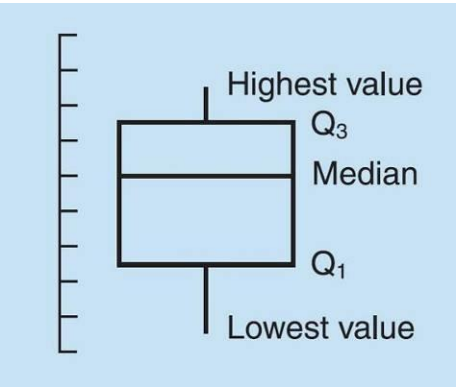
We will use these five numbers to create a graphic sketch of the data called a *box-and-whisker plot*. Box-and-whisker plots provide another useful technique from exploratory data analysis (EDA) for describing data.

Box-and-Whisker Plots

Procedure

HOW TO MAKE A BOX-AND-WHISKER PLOT

1. Draw a vertical scale to include the lowest and highest data values.
2. To the right of the scale, draw a box from Q_1 to Q_3.
3. Include a solid line through the box at the median level.
4. Draw vertical lines, called *whiskers*, from Q_1 to the lowest value and from Q_3 to the highest value.



Box-and-Whisker Plot

Figure 3-6

The next example demonstrates the process of making a box-and-whisker plot.

Example 10 – *Box-and-whisker plot*

Make a box-and-whisker plot showing the calories in vanilla-flavored ice cream bars. Use the plot to make observations about the distribution of calories.

Solution:

(a) We ordered the data (see Table 3-7) and found the values of the median, Q_1, and Q_3.

111	131	147	151	151	182
182	190	197	201	209	234
286	294	295	310	319	342
353	377	377	439		

Ordered Data

Table 3-7

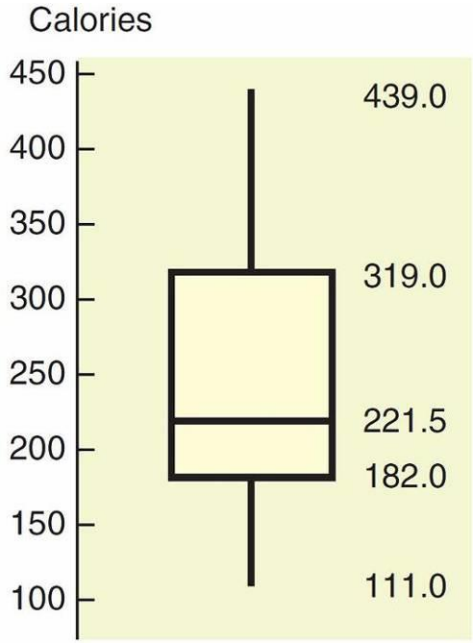
Example 10 – *Box-and-whisker plot* cont'd

From this previous work we have the following five-number summary:

low value = 111; Q_1 = 182; median = 221.5; Q_3 = 319; high value = 439

Example 10 – *Box-and-whisker plot* cont'd

(b) We select an appropriate vertical scale and make the plot (Figure 3-7).



Box-and-Whisker Plot for Calories in Vanilla-Flavored Ice Cream Bars

Figure 3-7

Example 10 – *Box-and-whisker plot* cont'd

(c) *Interpretation* A quick glance at the box-and-whisker plot reveals the following:

(i) The box tells us where the middle half of the data lies, so we see that half of the ice cream bars have between 182 and 319 calories, with an interquartile range of 137 calories.

(ii) The median is slightly closer to the lower part of the box. This means that the lower calorie counts are more concentrated. The calorie counts above the median are more spread out, indicating that the distribution is slightly skewed toward the higher values.

Example 10 – *Box-and-whisker plot* cont'd

(iii) The upper whisker is longer than the lower, which again emphasizes skewness toward the higher values.