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.

Median and percentiles

The sample median of a set of n measurements $x_1, \ldots, x_n$ is the middle value when the measurements are arranged from smallest to largest.

Calculating the Sample 100p-th Percentile

1. Order the data from smallest to largest.
2. Determine the product (*sample size*) $\times$ (*proportion*) $=$ np.

If np is not an integer, round it up to the next integer and find the corresponding ordered value.

If np is an integer, say k, calculate the average of the kth and $(k + 1)$st ordered values.

Sample Quartiles

Lower (first) quartile	Q_1 = 25th percentile
Second quartile (or median)	Q_2 = 50th percentile
Upper (third) quartile	Q_3 = 75th percentile

Example

Table 8 The Lengths of Long-Distance Phone Calls in Minutes

1.6	1.7	1.8	1.8	1.9	2.1	2.5	3.0	3.0	4.4
4.5	4.5	5.9	7.1	7.4	7.5	7.7	8.6	9.3	9.5
12.7	15.3	15.5	15.9	15.9	16.1	16.5	17.3	17.5	19.0
19.4	22.5	23.5	24.0	31.7	32.8	43.5	53.3		

To determine the first quartile, we take p=0.25 and calculate the product 38*0.25=9.5. Since 9.5 is not an integer, we take the next largest integer 10. We can see from the above data that 10^{th} ordered observation is 4.4. So, the first quartile (Q1)= 4.4.

For second quartile, take p=0.50, and 38*0.50= 19. Since this is an integer, we average the 19^{th} and 20^{th} observation from the ordered Data i.e (9.3+9.5)/2 = 9.4

4

Example

a. Determine the median of the following distribution.

58, 67, 60, 84, 93, 98, 100

b. Determine the median of the following distribution.

27, 19, 28, 18, 19, 35

c. Using the information provided in parts (a) and (b). Determine the first quartile (Q1), third quartile (Q3), and interquartile range (IQR) for the given data sets (Q3- Q1).

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*

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 =n*P=27*0.5=13.5 ~14th value

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

Similarly First quartile = $Q_1 = 0.33$

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

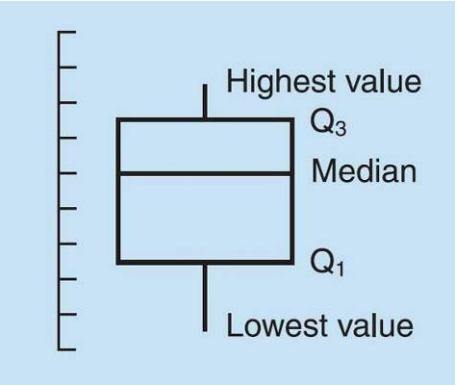
We will use these five numbers to create a graphic sketch of the data called a *box-and-whisker plot.* Box-and-whisker plots provide another useful technique from exploratory data analysis (EDA) for describing data.

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.

(a) Ordered the data and find the values of five number summaries

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

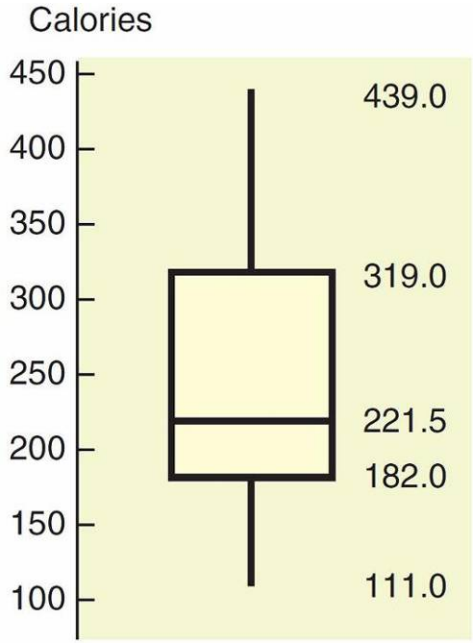
Example 10 – *Box-and-whisker plot* cont'd

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.

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

Box-Plot

Data:

3 7 2 10 3 1 8

4 4 3 2 4 5 9

6 9 5 2 8 7

1 2 2 2 3 3 3 4 4 4 5 5 6 7 7 8 8 9 9 10

Minimum = 1

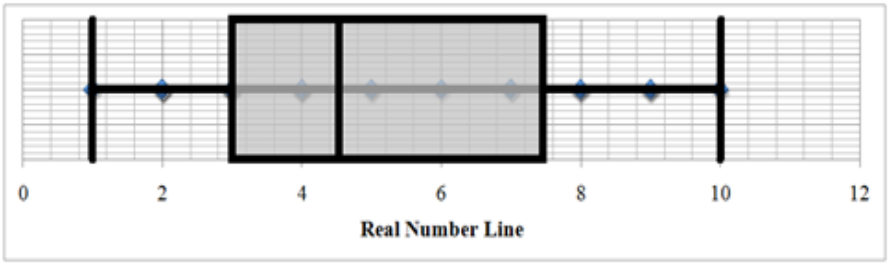
Maximum = 10

Median = 4.5

Q_1=3

Q_3=7.5

Software may apportion the values and give Q_3=7.25



Rules for Outliers

1. Under the assumption that the distribution is bell-shaped and symmetric. Any point further than three **standard deviations** from the **mean, $\mu \pm 3\sigma$**.

2. Any point further outside the interval between the **first quartile** minus one and a half times the **interquartile range** to the **third quartile** plus one and a half times the **interquartile range**,

$$\mathbf{(Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR)}.$$

Check if there is outlier exists in the above example

19

GUIDED EXERCISE 8 | BOX-AND-WHISKER PLOT

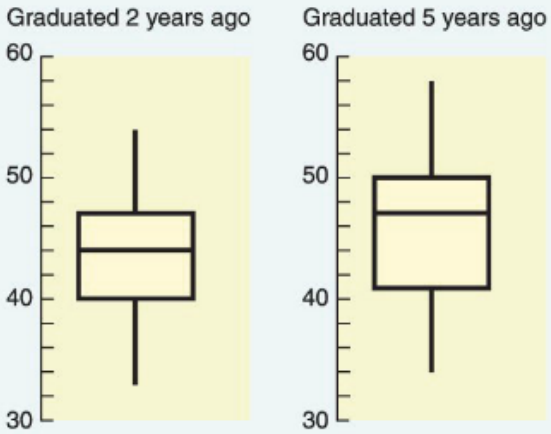
The Renata College Development Office sent salary surveys to alumni who graduated 2 and 5 years ago. The voluntary responses received are summarized in the box-and-whisker plots shown in Figure 3-8 on the next page.

(a) From Figure 3-8, estimate the median and extreme values of salaries of alumni graduating 2 years ago. In what range are the middle half of the salaries?

➡ The median seems to be about \$44,000. The extremes are about \$33,000 and \$54,000. The middle half of the salaries fall between \$40,000 and \$47,000.

FIGURE 3-8

Box-and-Whisker Plots for Alumni Salaries (in thousands of dollars)



(b) From Figure 3-8, estimate the median and the extreme values of salaries of alumni graduating 5 years ago. What is the location of the middle half of the salaries?

➡ The median seems to be \$47,000. The extremes are \$34,000 and \$58,000. The middle half of the data is enclosed by the box with low side at \$41,000 and high side at \$50,000.

(c) *Interpretation* Compare the two box-and-whisker plots and make comments about the salaries of alumni graduating 2 and 5 years ago.

➡ The salaries of the alumni graduating 5 years ago have a larger range. They begin slightly higher than and extend to levels about \$4000 above the salaries of those graduating 2 years ago. The middle half of the data is also more spread out, with higher boundaries and a higher median.

Example

Some data sets include values so high or so low that they seem to stand apart from the rest of the data. These data are called outliers. Outliers may represent data collection errors, data entry errors, or simply valid but unusual data values. It is important to identify outliers in the data set and examine the outliers carefully to determine if they are in error. One way to detect outliers is to use a box- and-whisker plot. Data values that fall beyond the limits

Lower limit: Q1- 1.5*(IQR)

Upper limit: Q3+ 1.5*(IQR)

Where IQR is the interquartile range, are suspended outliers.

Students from a statistics class were asked to record their heights in inches. The heights (as recorded) were:

65 72 68 64 60 55 73 71 52 63 61 74 69 67 74 50 4 75 67 62 66 80

64 65

Solution

Make a box-and whisker plot of the data. **(Low=4, Q1= 61.5, Median= 65.5, Q3= 71.5, high =80) (Draw the plot using these values**

Find the value of the interquartile range (IQR). **(10)**

Find the upper and lower limits. **(46.5, 86.5)**

Are there any data values below the lower limit? Above the upper limit? List any suspected outliers. What might be some explanations for the outliers? **(Yes, 4 is below the lower limit and is probably and error)**