3.4: The Five-Number Summary and Boxplots

Percentiles:

The kth percentile, denoted P_k, of a data set is the value such that k % of the data points are less than or equal to that value. The percentile rank of a score is the percent of scores equal to or below that score.

For example, a value is known as the 85th percentile if 85% of the data points are less than or equal to that score.

Example 1: Here are the 50 randomly generated scores from Example 4 in Section 3.3. Estimate the 70th percentile, 80th percentile and the 90th percentile.

37.48295	53.07996	54.94143	57.29676	60.95421	63.16013	66.48368
44.16628	53.20456	55.31494	57.37955	61.43636	63.3329	67.79641
47.40146	54.25092	55.90412	58.99277	61.91373	63.39574	67.85567
50.54246	54.41687	56.48669	59.10063	62.14886	63.61741	68.12883
51.77209	54.42467	56.64306	59.74812	62.52829	63.79043	68.23415
52.06366	54.87849	56.84053	60.00459	62.58302	63.93691	70.72309
53.05055	54.91449	57.00922	60.59386	63.15417	66.44211	73.3014
						87.41814

90th percentile

80th percentile

70th percentile

10% of 50 is 0.10(50) = 5.

So count off 5 values for the 90th percentile, another 5 for 80th percentile, etc.

Quartiles:

Quartiles are values that divide a data set into fourths. The 25th percentile, 50th percentile, and 75th percentile are often referred to as the first quartiles, second quartile, and third quartile.

Method 1 (Tukey's Method): Used in our book:

The second quartile, Q_2 , is the median M of the data set.

The first quartile, Q_1, is the median of the *bottom half of the data set.

The third quartile, Q_3, is the median of the *top half of the data set.

* If the data set has an odd number of data points, the median is included in both halves.

Method 2 (NOT Used in our book):

The second quartile, Q_2 , is the median M of the data set.

The first quartile, Q_1, is the median of the bottom half of the data set (the values less than M).

The third quartile, Q_3, is the median of the top half of the data set (the values greater than M).

Example 2: Calculate the quartiles for the data set $A = \{17,1,9,3,4,10,12,11,5,9,12,8,13,2,7\}$.

$M = Q_2$

1 2 3 4 5 7 8 9 9 10 11 12 12 13 17

$Q_1 = 4.5$

$Q_2 = M = 9$

$Q_3 = 11.5$

Median = Q_2

Bottom half: 1 2 3 4 5 7 8 9

$Q_1 = \frac{4+5}{2} = 4.5$

Top half

9 9 10 11 12 12 13 17

$Q_3 = \frac{11+12}{2} = 11.5$

Example 3: Calculate the quartiles for $B = \{2,3,3,4,5,5,6,7,9,9,9,10,11,12,12,13\}$.

$Q_2 = M = \frac{7+9}{2} = 8$

Bottom half: 2,3,3,4,5,5,6,7

$Q = \frac{4+5}{2} = 4.5$

Top half 9 9 9 10 11 12 12 13

$Q_3 = \frac{10+11}{2} = 10.5$

Example 4: Calculate the quartiles for $C = \{1,2,3,8,11,15,16,19,27,29,31,34,40,51,52,52,53\}$.

Example 5: Calculate the quartiles for $D = \{1,1,3,5,10,10,15,15,19,20,22,24,24,30,31,32,32,38\}$.

Definition: The *interquartile range*, denoted *IQR*, is the difference between the first and third quartiles.

$$IQR = Q_3 - Q_1$$

The *IQR* is the range of the middle 50% of the data set. The interquartile range is a measure of dispersion (how spread out the data are); the standard deviation, variance, and range of the data set are also measures of dispersion. The IQR is resistant to extreme values (outliers); the range and standard deviation are not resistant to extreme values.

An *outlier* is an extreme value (extremely low or extremely high, relative to other values in the data set).

One common definition for an outlier: A data point is considered an outlier (or a potential outlier) if it lies beyond these *fences*:

$$\text{Lower fence (lower limit)} = Q_1 - 1.5(IQR)$$
$$\text{Upper fence (upper limit)} = Q_3 + 1.5(IQR)$$

So, a data point x is an outlier if $x < Q_1 - 1.5(IQR)$ or if $x > Q_3 + 1.5(IQR)$.

Example 6: Using the definition above, find any outliers in these data sets.

a. $A = \{2,5,7,10,12,14,30\}$

bottom half / top half

$m = Q_2 = 10$

Bottom half: 2 5 7 10 $\quad Q_1 = \frac{5+7}{2} = 6$

Top half: 10 12 14 30 $\quad Q_3 = \frac{12+14}{2} = 13$

Quartiles: $Q_1 = 6$, $Q_2 = m = 10$, $Q_3 = 13$

Interquartile Range: $IQR = Q_3 - Q_1 = 13 - 6 = 7$

$1.5(IQR) = 1.5(7) = 10.5$

Lower fence: $Q_1 - 1.5(IQR) = 6 - 1.5(7) = 6 - 10.5 = -4.5$

Upper fence: $Q_3 + 1.5(IQR) = 13 + 1.5(7) = 13 + 10.5 = 23.5$

30 is beyond the upper fence, so 30 is an outlier

b. $B = \{2,14,16,19,23,24,30\}$

Example 7: Does the randomly generated data set in Example 1 contain any outliers?

Some researchers and statisticians consider a data point to be an extreme outlier if it lies beyond the two outer fences $Q_1 - 3(IQR)$ and $Q_3 + 3(IQR)$. Does the Example 1 data set contain extreme outliers?

The five-number summary:

We can get a fairly useful and descriptive picture of any data set from just 5 numbers: the minimum (smallest value), first quartile, second quartile (median), third quartile, and maximum (largest value).

Five-number summary:				
Minimum	Q_1	Q_2	Q_3	Maximum

Five-number summary for Example 6a

Min	Q_1	$Q_2 = M$	Q_3	Max
2	6	10	13	30

Boxplots:

A boxplot, or box-and-whisker plot, visually depicts these five numbers.

How to make a boxplot:

1. Determine the minimum, quartiles, and maximum of the data set.
2. Set up a horizontal scale, and draw a box that has Q_1 and Q_3 for endpoints, and a vertical line at Q_2 (the median). The length of the box is $IQR = Q_3 - Q_1$.
3. Calculate the upper and lower fences, and mark them on the graph:

   Lower fence = $Q_1 - 1.5(IQR)$
   Upper fence = $Q_3 + 1.5(IQR)$

4. Draw a line from Q_1 to the smallest data point that is larger than the lower fence. Draw a line from Q_3 to the largest data point that is smaller than the upper fence.
5. Use an asterisk to mark any data values that lie outside the fences.

Example 8: Construct a box plot for the data set.

3, 4, 4, 5, 5, 5, 6, 6, 7, 7, 7, 7, 8, 8, 9, 11

Example 9: Construct a box plot for the data set.

25, 1, 5, 3, 7, 14, 12, 10, 5, 9, 12, 4, 6, 13, 2, 8

1 2 3 4 5 5 6 7 8 9 10 12 12 13 14 25

$Q_1 = \frac{4+5}{2} = 4.5$ $\quad Q_2 = m = \frac{7+8}{2} = 7.5$ $\quad Q_3 = \frac{12+12}{2} = 12$

Min = 1
$Q_1 = 4.5$
$Q_2 = m = 7.5$
$Q_3 = 12$
Max = 25

$IQR = Q_3 - Q_1 = 12 - 4.5 = 7.5$

Lower fence: $Q_1 - 1.5(IQR) = 4.5 - 1.5(7.5) = -6.75$

Upper fence: $Q_3 + 1.5(IQR) = 12 + 1.5(7.5) = 23.25$

So 25 is an outlier