Ch 2.3 and 2.4 Percentile, Boxplot and Outliers

Percentile and Quartiles

Percentile: are measures of location. Denoted by $P_1, P_2, \ldots P_{99}$ which divide a set of data into 100 groups with about 1% of the values in each group.

If x is at 90^{th} percentile, means 90% of all data are less than x. Note, percentile is not the same as percentage.

Quartiles: (Q_1, Q_2, Q_3)

Quartiles are measures of location, which divide a set of data into four groups with about 25% of the values in each group.

Q_1 – First quartile or P_{25}. It separates the bottom 25% of value from the top 75%.

Q_2 - Second quartile or P_{50} or median. It separates the bottom 50% of values from the top 50%.

Q_3 – Third quartile or P_{75}. It separates the bottom 75% of values from the top 25%.

Five-number-summary, IQR and Boxplot:

Five-number-summary are:

Mininum, Q_1, Median, Q_3 and Maximum divides the data into four groups of 25% each.

$IQR = Q_3 - Q_1$ (Inter-quantile Range)

The **interquartile range (IQR)** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

A **boxplot** shows graphical image of concentration of data. A boxplot is constructed using 5-number summary with Q1, median and Q3 in a box containing 50% of all data. It gives good distribution of data in 25%, 50% and 75%.

- Maximum and Minimum values are extended as whiskers at the two ends of the box.

Find 5-number-summary and boxplot by Statdisk

- Enter data in a column in Statdisk.

- Select Data, Explore data, descriptive statistic, select column and Click evaluate.

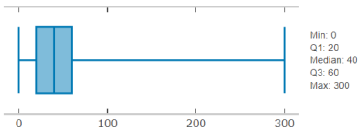
Ex1. The time(in min.) a sample of 15 student spent on exercising daily is given:

0, 40, 60, 30, 60, 10, 46, 30, 300, 90, 30, 120, 60, 0, 20

a) Find the 5-number summary and sketch a boxplot.

Use statdisk, data/explore data/select column data/evaluate, five number summary and boxplot will show.

Minimum:	0
1st Quartile:	20.00000
2nd Quartile:	40.00000
3rd Quartile:	60.00000
Maximum:	300



b) What percent of student exercise from 0 to 60 min?

Because 60 is Q3, so 75% of all student exercise from 0 to 60 min.

c) What percent of student exercise between 20 to 60 min?

Because 20 is Q1, 60 is Q3, so 50% of all students exercise from 20 to 60 min.

Answer: Use Statdisk:

a) Min = 0, Q1=20, Med=40, Q3=60, Max = 300

b) Since Q3 = 60, hence 75% of students exercise from 0 to 60 min.

c) Since Q1 = 20 and Q3=60, Hence 50% of students exercise from 20 to 60 min.

Outliers and IQR

IQR is used to determine potential **outliers**.

1. Find the quartiles Q_1, Q_2, and Q_3.
2. Find the interquartile range (IQR), where IQR $= Q_3 - Q_1$.
3. Evaluate $1.5 \times$ IQR.
4. **In a modified boxplot, a data value is an *outlier* if it is**
   above Q_3, by an amount greater than $1.5 \times$ IQR
   or below Q_1, by an amount greater than $1.5 \times$ IQR

Ex1. If Q1 = 34, Q3 = 70, find the lower fence and upper fence for an outlier.

IQR = 70 - 34 = 36

lower fence is 34 - 1.5(36) = -20, upper fence = 70 + 1.5(36) = 124

Value between -20 and 124 are not outliers, values outside the range are outliers.

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

C) Modified boxplot and outliers:

A modified boxplot can be graphed to show outliers without calculating IQR and applying the Q1-1.5IQR, Q3+1.5IQR. Outliers are shown as markers in the boxplot.

- use Statdisk, click data , Boxplot,

- Select the column of data, click **modified boxplot**. The outlier will be shown as marker at the lowest or highest end of the boxplot.

- If there are no markers, there is no outliers in the dataset.

- To find the values of the outlier, sort the data. The outliers will be at the top and end of the sorted data.

Ex2. Determine if outliers exist in the exercise time from 15 students.

0, 40, 60, 30, 60, 10, 46, 30, 300, 90, 30, 120, 60, 0, 20

By calculation:

Since Q1 = 20, Q3 = 60, So IQR = 60 – 20 = 40

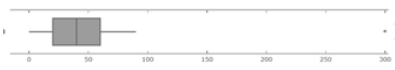
Lower fence = Q1 – 1.5(IQR) = 20 – 1.5(40) = -40

upper fence = Q3 + 1.5(IQR) = 60 + 1.5(40) = 120

Values lower than -40 and higher than 120 is an outlier. So the value 300 is an outlier.

Graph a modified boxplot to identify outliers.

Use Statdisk, Data/Boxplot/select Modified boxplot.



There is one outlier in the high end of the data. To find the outlier, sort the data and locate the highest value.

Use Statdisk, Sort, one column, select the column containing the data. The last data (300) is the outlier.
