Probability and Statistics

Topic 5 - The Five-Number Summary and Boxplots

Aamir Alaud Din, PhD

October 12, 2023

TABLE OF CONTENTS

TABLE OF CONTENTS

RECAP

- The z-score of a data point is the distance of the data point from the mean in terms of number of standard deviations.
- The population z-score is given by the formula

$$z = \frac{x - \mu}{\sigma}$$

- The sample z-score is given by the formula

$$z = \frac{x - \bar{x}}{s}$$

- The z-score can be used to compare the data points from two different data sets provided the mean and standard deviations for both data sets are known.
- The percentile divides the data set into 100 equal parts.

RECAP

- The kth percentile of a data set means k percent of the observations are less than or equal to the value.
- Quartiles divide the data set into four equal parts.
- First quartile Q_1 divides the bottom 25% data from the top 75% data.
- Second quartile Q_2 divides the bottom 50% data from the top 50% data which is in fact the median of the data set.
- Third quartile Q_3 divides the bottom 75% data from the top 25% data.
- Interquartile range is the range of the middle 50% of the observations in a data set and is given by the formula

$$\text{IQR} = Q_3 - Q_1$$

RECAP

- Outliers are the extreme observations in the data set.
- Outliers may or may not be a part of the data set.
- If an extreme observation is a typo mistake, for example, it must be removed from the data set.
- If an extreme observation, no matter how much extreme, is an observation from the sample, it must be kept in the data set and included in computations.
- The observations below the lower fence (shown below) and above the upper fence (also shown below) must be checked for outliers.

$$\text{Lower Fence} = Q_1 - 1.5(\text{IQR}) \qquad \text{and} \qquad \text{Upper Fence} = Q_3 + 1.5(\text{IQR})$$

TABLE OF CONTENTS

OBJECTIVES

After learning this topic and studying, you should be able to:

1. Compute the five-number summary
2. Draw and interpret boxplots

TABLE OF CONTENTS

THE WHY SECTION

- So far, we studied the measures of
  1. Central Tendency
  2. Dispersion
  3. Position
  4. Outliers
- They don't give a complete picture of the data.
- They also give no idea about the shape of the distribution.
- Five-number summary and boxplots give an idea of the above two points.

TABLE OF CONTENTS

FIVE-NUMBER SUMMARY

- Remember that the median is a measure of central tendency that divides the lower 50% of the data from the upper 50%.
- It is resistant to extreme values and is the preferred measure of central tendency when data are skewed right or left.
- The three measures of dispersion (range, standard deviation, and variance) are not resistant to extreme values.
- However, the interquartile range, $Q_3 - Q_1$, the difference between the 75th and 25th percentiles, is resistant.
- It is interpreted as the range of the middle 50% of the data.
- However, the median, Q_1, and Q_3 do not provide information about the extremes of the data, the smallest and largest values in the data set.

FIVE-NUMBER SUMMARY

- The five-number summary of a set of data consists of the smallest data value, Q_1, the median, Q_3, and the largest data value.
- We organize, the five-number summary as follows:

Five-Number Summary

MINIMUM Q_1 M Q_3 MAXIMUM

EXAMPLE 1

The data shown in Table 17 show the finishing times (in minutes) of the men in the 60- to 64-year-old age group in a 5-kilometer race. Determine the five-number summary of the data.

Table 17

19.95	23.25	23.32	25.55	25.83	26.28	42.47
28.58	28.72	30.18	30.35	30.95	32.13	49.17
33.23	33.53	36.68	37.05	37.43	41.42	54.63

TABLE OF CONTENTS

BOXPLOTS

- The five-number summary can be used to create another graph, called the boxplot.

Drawing a Boxplot

1. Determine the lower and upper fences:

$$\text{Lower Fence} = Q_1 - 1.5(\text{IQR}) \quad \text{and} \quad \text{Upper Fence} = Q_3 + 1.5(\text{IQR})$$

   where, $\text{IQR} = Q_3 - Q_1$

2. Draw a number line long enough to include the maximum and minimum values. Insert vertical lines at Q_1, M, and Q_3 . Enclose these vertical lines in a box.

3. Label the lower and upper fences.

BOXPLOTS

Drawing a Boxplot

4. Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence. These lines are called whiskers.

5. Any data values less than the lower fence or greater than the upper fence are outliers and are marked with an asterisk (*).

EXAMPLE 3

Use the results from Example 1 to construct a boxplot of the finishing times of the men in the 60- to 64-year-old age group.

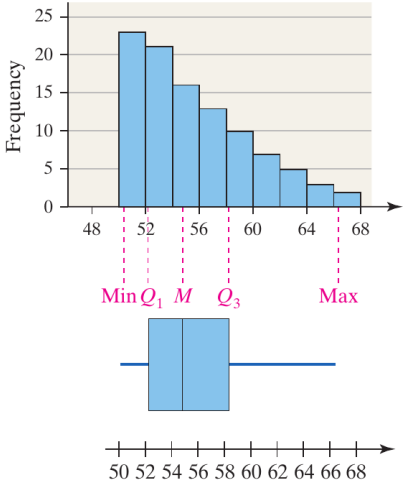
Using Boxplot and Quartiles to Describe the Shape of a Distribution

- Figure below shows three histograms and their corresponding boxplots with the five-number summary labeled. Notice the following from the figure.
- Figure (a), the histogram shows the distribution is skewed right.
- Notice that the median is left of center in the box, which means the distance from M to Q_1 is less than the distance from M to Q_3.
- In addition, the right whisker is longer than the left whisker.
- Finally, the distance from the median to the minimum value in the data set is less than the distance from the median to the maximum value in the data set.

BOXPLOTS

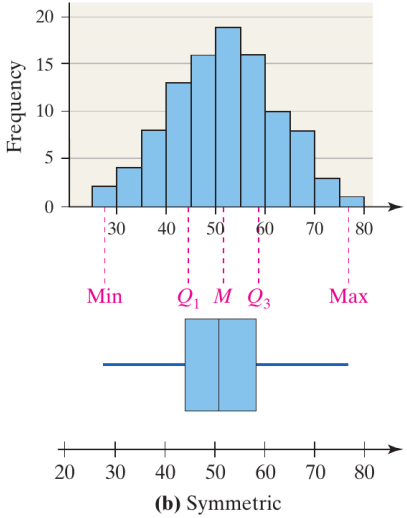
- In Figure (b), the histogram shows the distribution is symmetric.
- Notice that the median is in the center of the box, so the distance from M to Q_1 is the same as the distance from M to Q_3.
- In addition, the left and right whiskers are roughly the same length.
- Finally, the distance from the median to the minimum value in the data set is the same as the distance from the median to the maximum value in the data set.
- In Figure (c), the histogram shows the distribution is skewed left.
- Notice that the median is right of center in the box, so the distance from M to Q_1 is more than the distance from M to Q_3.

BOXPLOTS



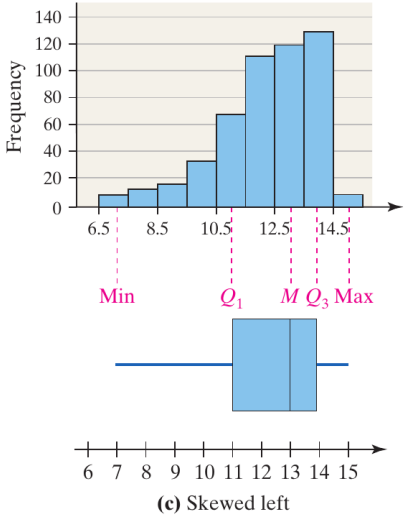
(a) Skewed right

BOXPLOTS



(b) Symmetric

BOXPLOTS



(c) Skewed left

BOXPLOTS

- In addition, the left whisker is longer than the right whisker.
- Finally, the distance from the median to the minimum value in the data set is more than the distance from the median to the maximum value in the data set.
- The guidelines given above are just that—guidelines.
- Judging the shape of a distribution is a subjective practice.
- The boxplot in Figure below suggests that the distribution is skewed right, since the right whisker is longer than the left whisker and the median is left of center in the box.
- We can also assess the shape using the quartiles.

BOXPLOTS

- The distance from M to Q_1 is 4.89 (=30.95-26.06), while distance from M to Q_3 is 6.29 (=37.24-30.95).
- Also, the distance from M to the minimum value is 11 (=30.95-19.95), while the distance from M to the maximum value is 23.68(=54.63-30.95).

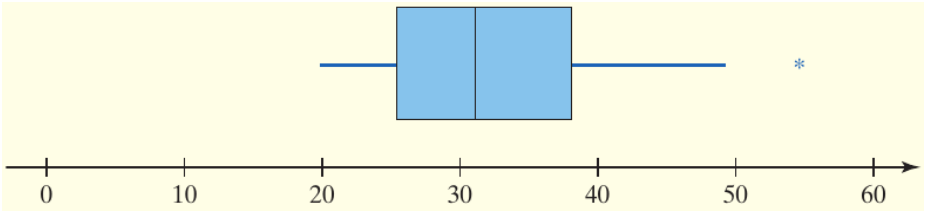


TABLE OF CONTENTS

SUMMARY

- The five-number summary of a data set includes minimum, Q_1, $Q_2 = M$, Q_3, and maximum.
- The five number summary gives the idea of the shape of distribution.
- The graphical representation of five-number summary is the box plot.
- The left, middle, and right lines of the box represent Q_1 , $Q_2 = M$, and Q_3.
- The extension lines of box represent the minimum and maximum values.
- The bounds and asterisk symbol represent the lower and upper fences and the outliers, respectively.


Thank You!