Introduction to Stochastics
Exercise 3

1 Analyzing Travel Times

Commute times (in minutes): 25, 30, 28, 32, 30, 60, 27, 31

- Find the mean, median, and mode!
- What is the range?
- Are there any outliers? Justify your answer using the IQR method!

1.1 Solution

Ordered times: 25, 27, 28, 30, 30, 31, 32, 60

$$\text{Mean} = \frac{25 + 27 + 28 + 30 + 30 + 31 + 31 + 60}{8} = \frac{263}{8} = 32.875$$

Since there are 8 values (even number), the median is the average of the 4th and 5th values:

$$\text{Median} = \frac{30 + 30}{2} = 30$$

The value that occurs most often is:

$$\text{Mode} = 30$$

$$\text{Range} = \text{Maximum} - \text{Minimum} = 60 - 25 = 35$$

Outliers (IQR Method):
First, find the quartiles from the ordered data:

$$\begin{aligned} Q_1 &= \text{Median of } \{25, 27, 28, 30\} = \frac{27 + 28}{2} = 27.5 \\ Q_3 &= \text{Median of } \{30, 31, 32, 60\} = \frac{31 + 32}{2} = 31.5 \\ IQR &= Q_3 - Q_1 = 31.5 - 27.5 = 4 \end{aligned}$$

Lower fence:

$$Q_1 - 1.5 \times IQR = 27.5 - 6 = 21.5$$

Upper fence:

$$Q_3 + 1.5 \times IQR = 31.5 + 6 = 37.5$$

Any value below 21.5 or above 37.5 is an outlier.
Since 60 is greater than 37.5, it is an outlier.

$$\text{Outlier: } 60$$

2 Interpreting a Boxplot

Below is a description of a boxplot (not drawn)
Min: 5, Q1: 8, Median: 10, Q3: 13, Max: 20

- What is the IQR?
- Is the distribution likely right-skewed, left-skewed, or symmetric?
- If a value of 30 were added, how would that affect the plot?

2.1 Solution

$$IQR = Q_3 - Q_1 = 13 - 8 = 5$$

The distribution is probably right-skewed.

The maximum would increase. In this case the upper fence would be calculated using Q_3+1.5*IQR (20.5), because the value 30 would be an outlier. The outlier would be symbolized using a dot.

3 Draw the boxplot with the following data.

An exam was conducted for students in two different classes. The results are summarized using five-point statistics (minimum, Q_1, median, Q_3, maximum). Using these summaries, draw a separate boxplot for each class.
Class A: (60, 70, 75, 80, 85), Class B: (40, 60, 70, 90, 100)

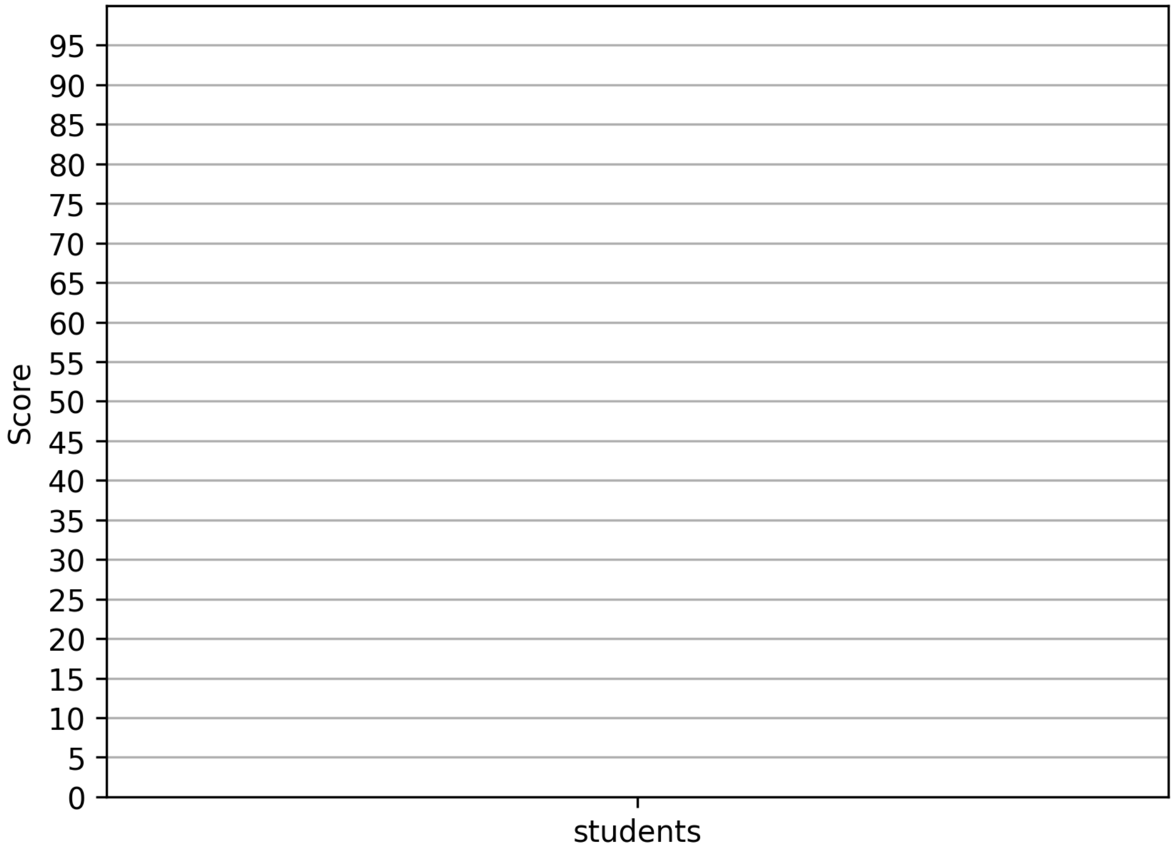

Figure 1: Empty Boxplots.

3.1 Solutions

Each boxplot visualizes the five-number summary of a data set:

- The box extends from the first quartile (Q_1) to the third quartile (Q_3).
- The median is shown as a horizontal line inside the box.
- The ”whiskers” typically extend to the smallest and largest data points that are not considered outliers.
- A data point is considered an outlier if it lies below the lower fence (Q_1 $-1.5\times$ IQR) or above the upper fence (Q_3 $+1.5\times$ IQR), where IQR is the interquartile range: $Q_3 - Q_1$.

In this case, no values lie beyond the fences for either class, so the whiskers simply extend to the minimum and maximum values. Therefore, both boxplots reflect the full range of scores without any outliers.

Note: If the maximum (as indicated by the 5-point summaries) would be an outlier, it wouldn’t be possible to draw the boxplot, because we wouldn’t know if there are any outliers that are smaller than the maximum.

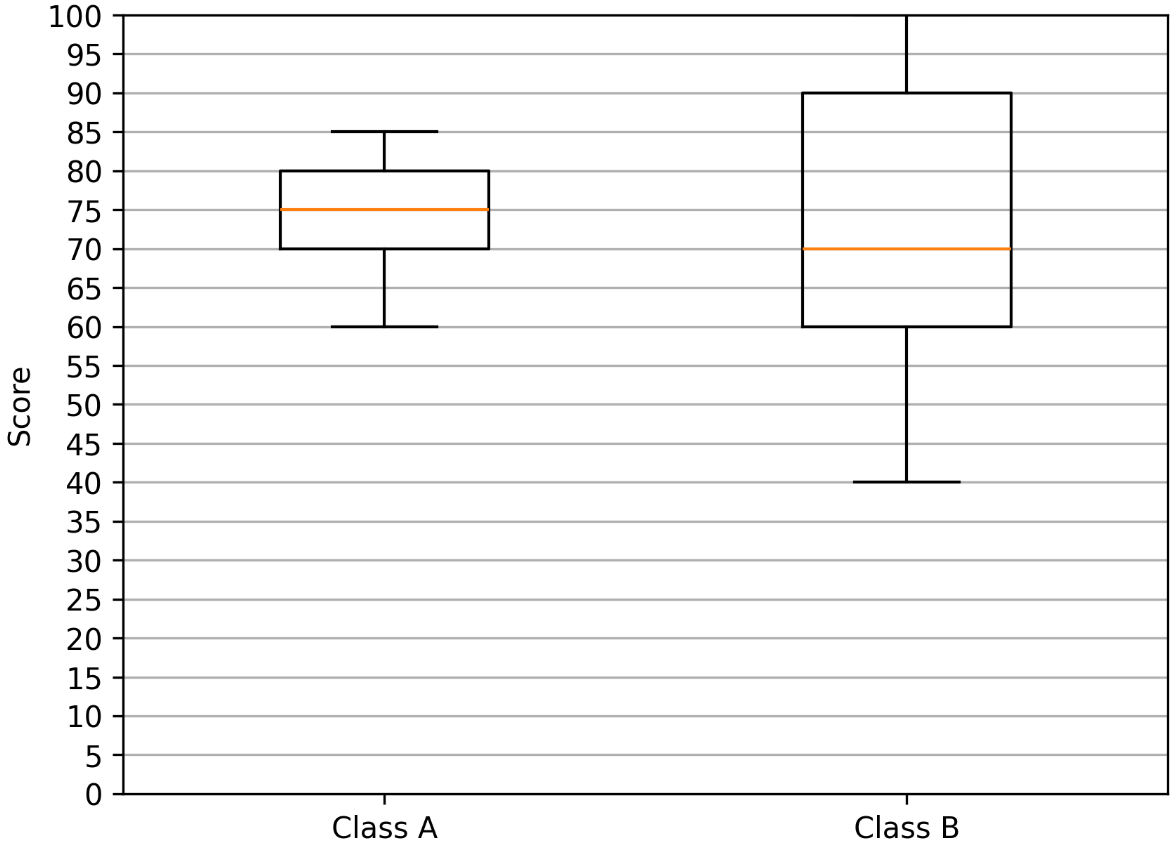

Figure 2: Boxplots.

4 Comparing Two Data Sets

Consider the following data sets of ages:

- **Data Set A (ages):** 10, 12, 11, 12, 10, 13
- **Data Set B (ages):** 10, 14, 16, 8, 20, 10

- Find the range and IQR for each set!
- Which set is more spread out?
- Which would you say has more consistent data, and why?

4.1 Solution

- **Range and IQR:**

  Data Set A:

  - Minimum = 10, Maximum = 13
  - **Range:**

$$\text{Range}_A = 13 - 10 = 3$$

  To find the IQR, first, calculate the quartiles:

    * Ordered Data: 10, 10, 11, 12, 12, 13
    * Q_1 = Median of {10, 10, 11} = 10
    * Q_3 = Median of {12, 12, 13} = 12
    * **IQR:**

$$\text{IQR}_A = 12 - 10 = 2$$

Data Set B:

  - Minimum = 8, Maximum = 20
  - **Range:**

$$\text{Range}_B = 20 - 8 = 12$$

  For the IQR, calculate the quartiles:

    * Ordered Data: 8, 10, 10, 14, 16, 20
    * Q_1 = Median of {8, 10, 10} = 10
    * Q_3 = Median of {14, 16, 20} = 16
    * **IQR:**

$$\text{IQR}_B = 16 - 10 = 6$$

- **Which set is more spread out?**

  Data Set B has a larger range (12 compared to 3) and a larger interquartile range (6 compared to 2), indicating that it is more spread out than Data Set A.

- **Which set has more consistent data, and why?**

  Data Set A has more consistent data. The smaller range (3) and smaller IQR (2) indicate that the values are clustered more closely together. Data Set B, on the other hand, has a wider range and IQR, which suggests a greater variability and less consistency in the data.

5 Number of Books Read in a Month by 12 Students

Number of books read in a month by 12 students: 3, 4, 2, 5, 3, 4, 4, 3, 3, 2, 6, 3

- What is the mode and what does it tell us?
- Find the median and range!
- Imagine a new student reads 15 books. How does that affect the mean? Is it still a useful measure?

5.1 Solution

- **Mode:**

  The mode is the number that appears most frequently in the data set.

  Ordered data: 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 6

  The number **3** appears most frequently (6 times). Therefore, the mode is:

  $$\text{Mode} = 3$$

  The mode tells us that most students read 3 books in a month, indicating this as the most common reading behavior among the group.

- **Median and Range:**

  To find the median, first, we order the data:

  $$\text{Ordered Data: } 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 6$$

  Since there are 12 data points (even number), the median is the average of the 6th and 7th values:

  $$\text{Median} = \frac{3+3}{2} = 3$$

  To calculate the range:

  $$\text{Range} = \text{Maximum} - \text{Minimum} = 6 - 2 = 4$$

- **Impact of a New Student Reading 15 Books:**

  If a new student reads 15 books, the updated data set will be:

  $$2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 6, 15$$

  To find the new mean, we first calculate the sum of the data:

  $$\text{Sum of new data} = 2 + 2 + 3 + 3 + 3 + 3 + 3 + 4 + 4 + 4 + 5 + 6 + 15 = 57$$

  The new mean is:

  $$\text{Mean} = \frac{57}{13} \approx 4.4$$

  The original mean was:

  $$\text{Original Mean} = \frac{42}{12} = 3.5$$

  The addition of the student reading 15 books increases the mean (from 3.5 to approximately 4.4).

  This shows that the mean can be heavily influenced by extreme values (outliers), making it less representative of the typical data in this case. While the mean gives us an overall average, it may no longer accurately reflect the typical number of books read by most students due to the influence of the outlier.