

Statistics

Lecture 20: | Boxplot

Prof. Dr. Mostafa Elhosseini

<https://youtube.com/drmelhosseini>

| What is a Boxplot

- For some distributions/datasets, you will find that you **need more information than** the measures of **central tendency** (median, mean, and mode)
- In **descriptive statistics**, a **box plot** or **boxplot** (also known as **box and whisker plot**) is a type of chart often used in explanatory data analysis (EDA)
- Box plots show the **five-number summary of a set of data**: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score

| Why Boxplot?

- You need to have information on the variability or dispersion of the data.
- A **boxplot** is a graph that gives you a **good indication of how the values in the data are spread out**.
- It can be used to **determine the regions in which the sample values are more densely crowded** and the regions in which they are **more sparse**
- Although boxplots may seem primitive in comparison to a histogram or density plot, they **have the advantage of taking up less space**, which is **useful when comparing** distributions between many groups or datasets.

| Interquartile Range (IQR)

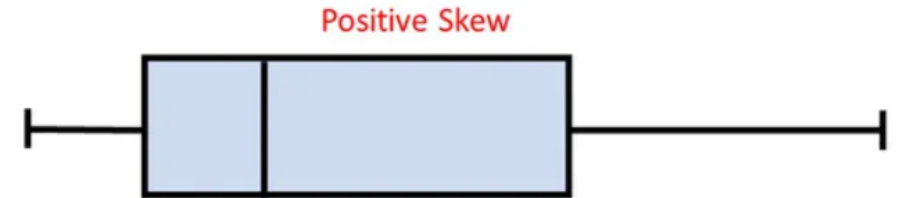
- The interquartile range is therefore the distance needed to span the middle half of the data
- Any point that is **more than 1.5 IQR above the third quartile**, or more than 1.5 IQR below the first quartile, is considered an **outlier**.
- Some texts define a point that is **more than 3 IQR from the first or third quartile** as an **extreme outlier**

| How to Plot?

- A **boxplot** is a standardized way of **displaying the distribution** of data based on a **five number summary**
 - Minimum: $Q1 - 1.5 * IQR$
 - First quartile (Q1),
 - Median (Q2),
 - Third quartile (Q3), and
 - Maximum: $Q3 + 1.5 * IQR$
- It can tell you about your outliers and what their values are.
- It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

| How to Judge Skewness using Boxplot

- When the **median is closer to the bottom** of the box, and if the **whisker is shorter on the lower** end of the box, then the distribution is **positively skewed (skewed right)**.
- When the **median is closer to the top** of the box, and if the **whisker is shorter on the upper** end of the box, then the distribution is **negatively skewed (skewed left)**.



| Variability - Dispersion

- In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed
- The smallest value and largest value are found at the end of whiskers and are useful for providing a visual indicator regarding the spread of scores (range)
- The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile (e.g. $Q3 - Q1$).

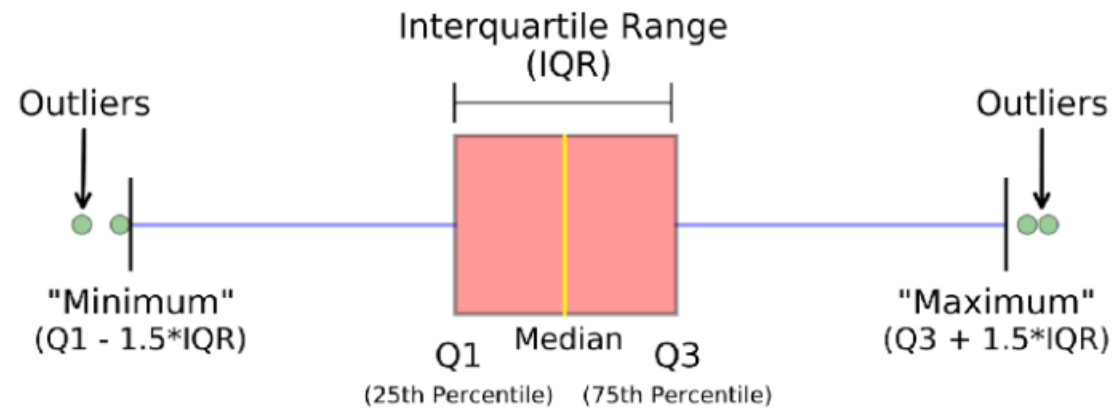
| Outliers

- An **outlier** is an observation that is numerically **distant from the rest of the data**
- When reviewing a box plot, an **outlier** is defined as a data point that is **located outside the whiskers of the box plot**
- Outside 1.5 times the IQR above the Q3 and below Q1 ($Q1 - 1.5 * IQR$ or $Q3 + 1.5 * IQR$)

| Steps in the Construction of a Boxplot

- Compute the median and the first and third quartiles of the sample. Indicate these with horizontal lines. Draw vertical lines to complete the box.
- Find the largest sample value that is no more than 1.5 IQR above the third quartile, and the smallest sample value that is no more than 1.5 IQR below the first quartile. Extend vertical lines (whiskers) from the quartile lines to these points.
- Points more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile, are designated as outliers. Plot each outlier individually.

| Steps in the Construction of a Boxplot



| References

- Understanding Boxplots <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>
- What does a box plot tell you?
<https://www.simplypsychology.org/boxplots.html>
- Statistics for Engineers and Scientists, Fourth Edition, William Navidi
Colorado School of Mines
- Khan Academy
<https://www.khanacademy.org/math/cc-sixth-grade-math/cc-6th-data-statistics/cc-6th-box-whisker-plots/v/constructing-a-box-and-whisker-plot>