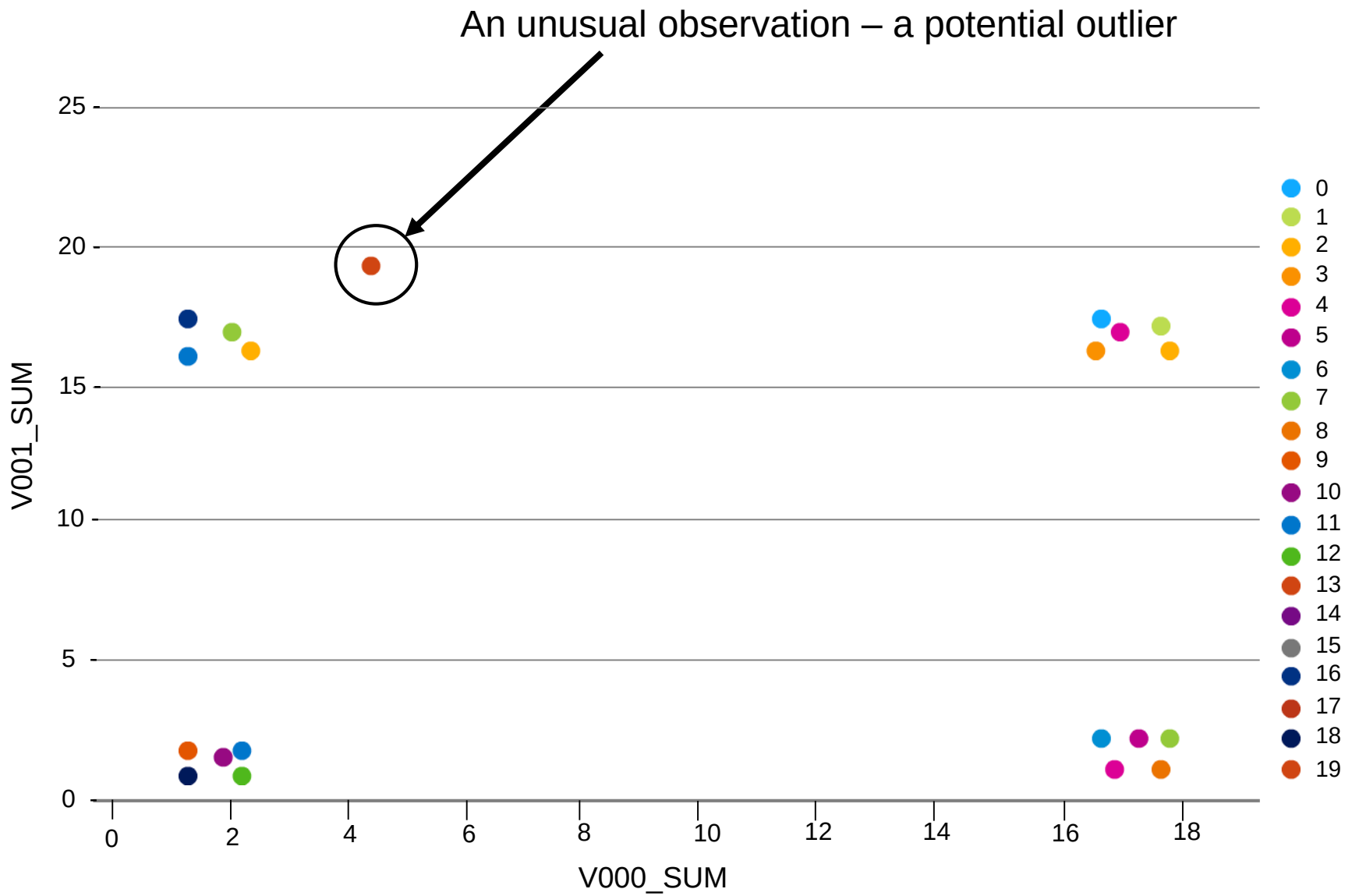
Week 2: Descriptive Statistics

Unit 6: Outliers

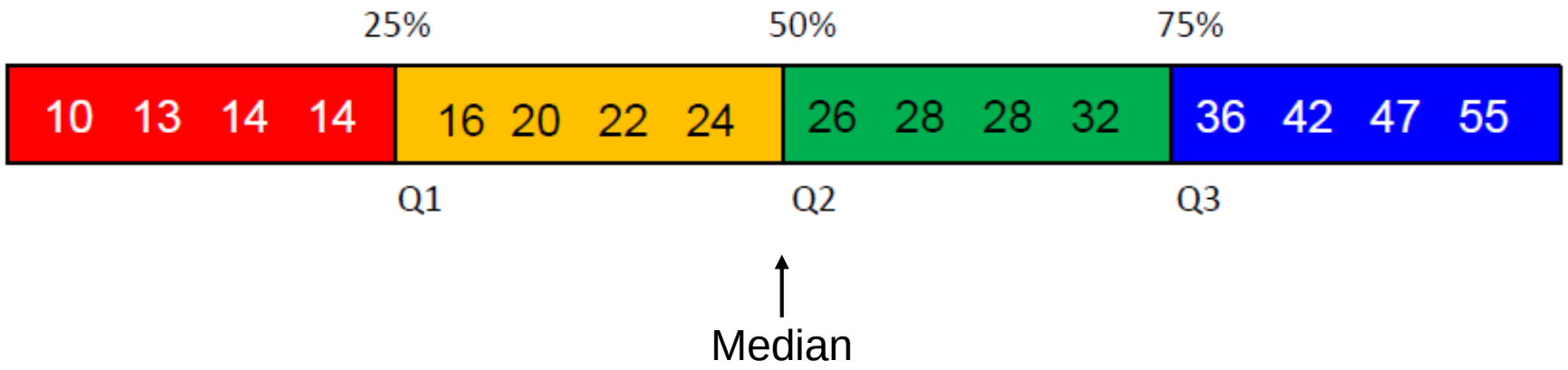
Outliers

Introduction

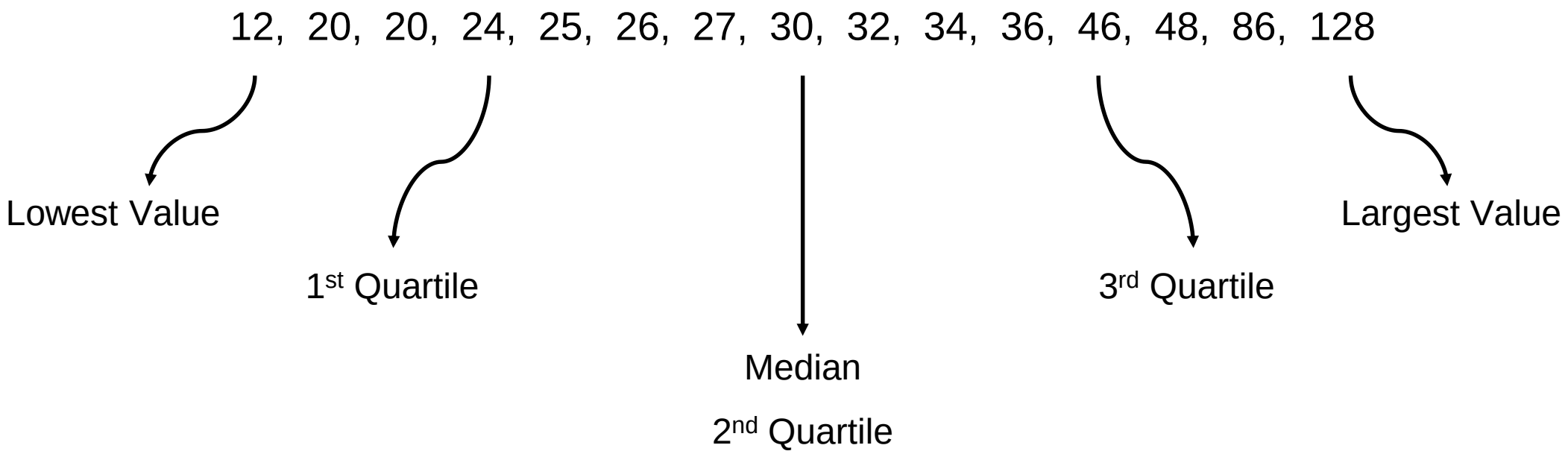


Understanding quartiles

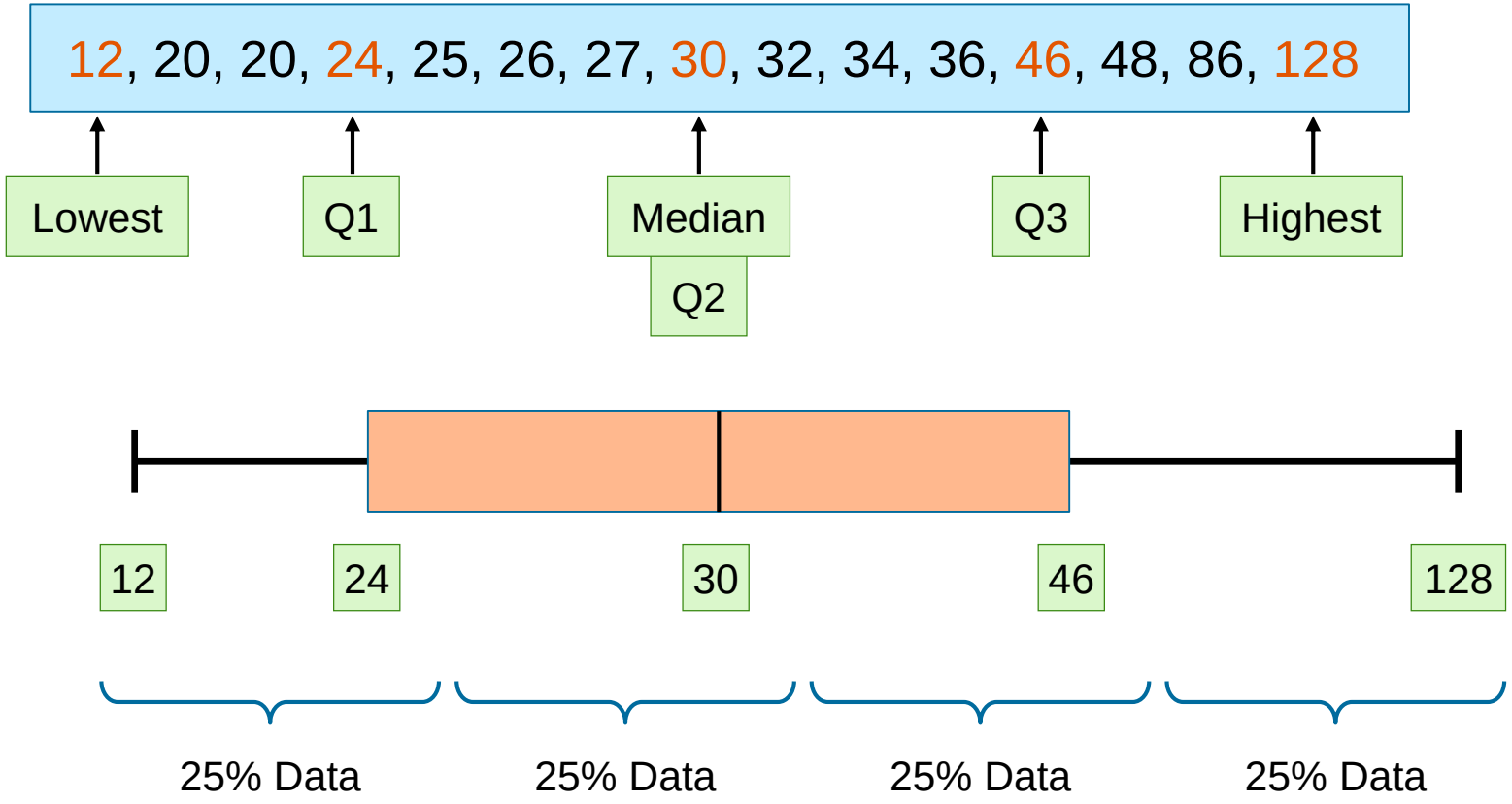
Quartile	Percentile
1st Quartile	25th Percentile
2nd Quartile	50th Percentile
3rd Quartile	75th Percentile



Quartiles

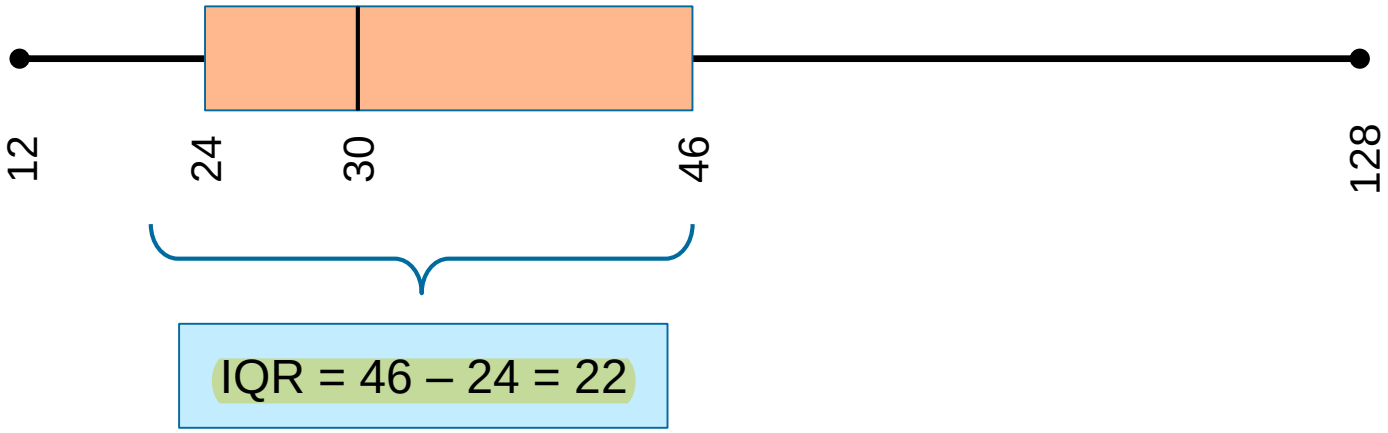


Detecting outliers using a box plot



Interquartile range (IQR)

Interquartile Range = Q3 – Q1



How to identify outlier and extreme value

- Fences are usually found with the following formulas:

  Upper fence = Q3 + (1.5 * IQR)

  Lower fence = Q1 – (1.5 * IQR)

- Sometimes you will see reference to “inner fences” and “outer fences”.

  Lower inner fence: Q1 – (1.5 * IQR)

  Upper inner fence: Q3 + (1.5 * IQR)

  Lower outer fence: Q1 – (3 * IQR)

  Upper outer fence: Q3 + (3 * IQR)

Outlier

Left Inner Fence	= Q1 - 1.5 (IQR)	➔	24 – 1.5 (22)	= -9
Right Inner Fence	= Q3 + 1.5 (IQR)	➔	46 + 1.5 (22)	= 79

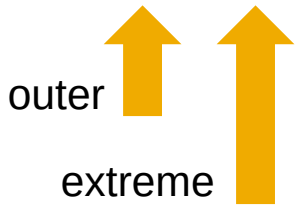
A value that is less than -9 or greater than 79 is an outlier

Outlier

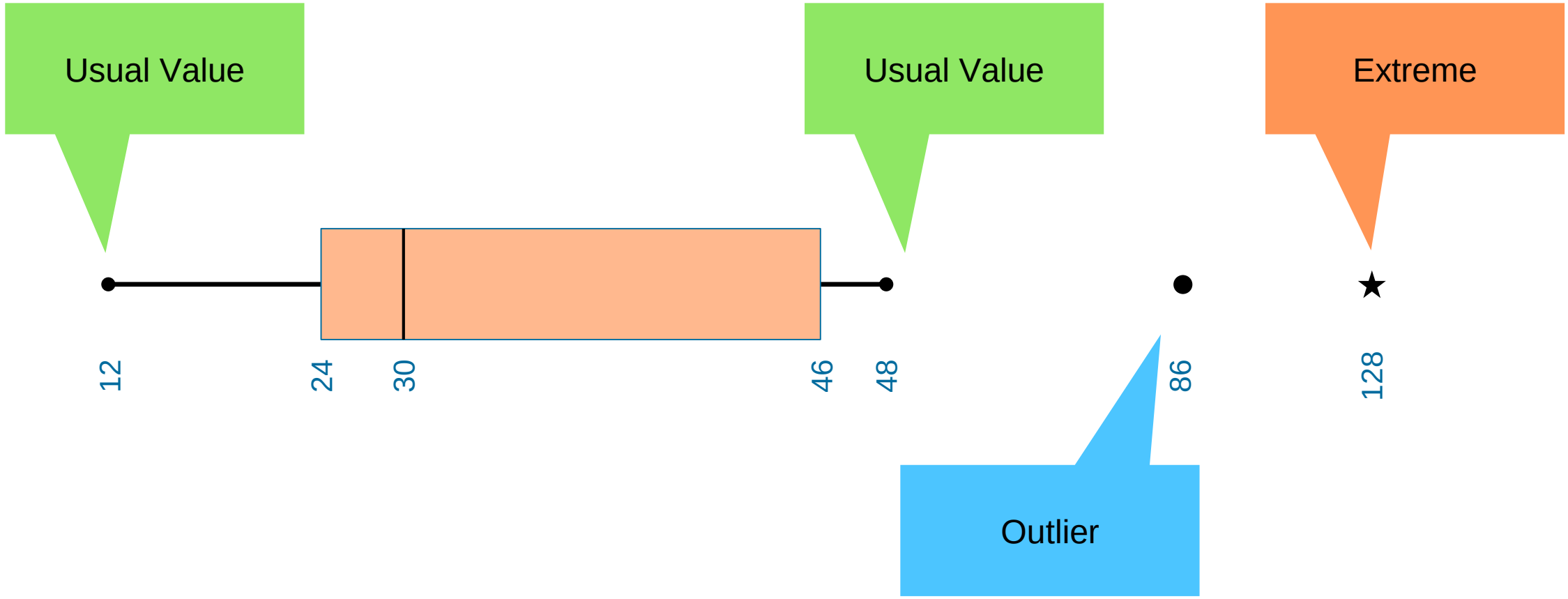
Left Outer Fence	= Q1 - 3.0 (IQR)	➔	24 – 3.0 (22)	= -42
Right Outer Fence	= Q3 + 3.0 (IQR)	➔	46 + 3.0 (22)	= 112

A value that is less than -42 or greater than 112 is an extreme value

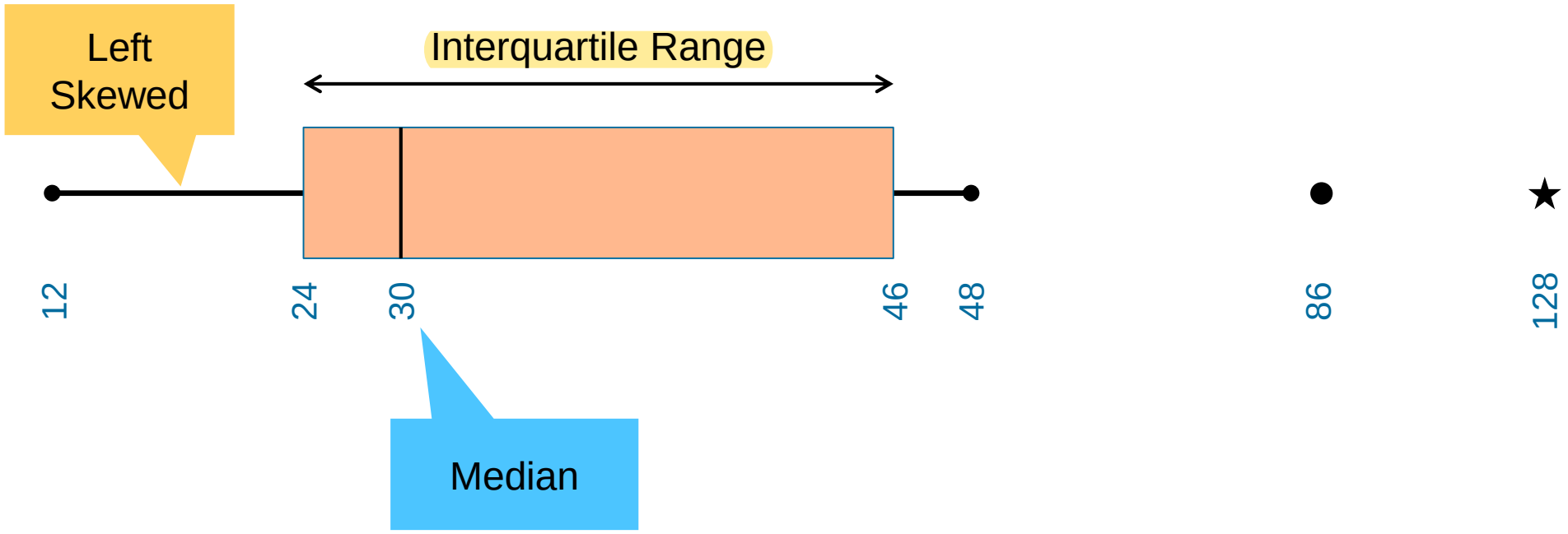
12, 20, 20, 24, 25, 26, 27, 30, 32, 34, 36, 46, 48, 86, 128



Outlier and extreme value

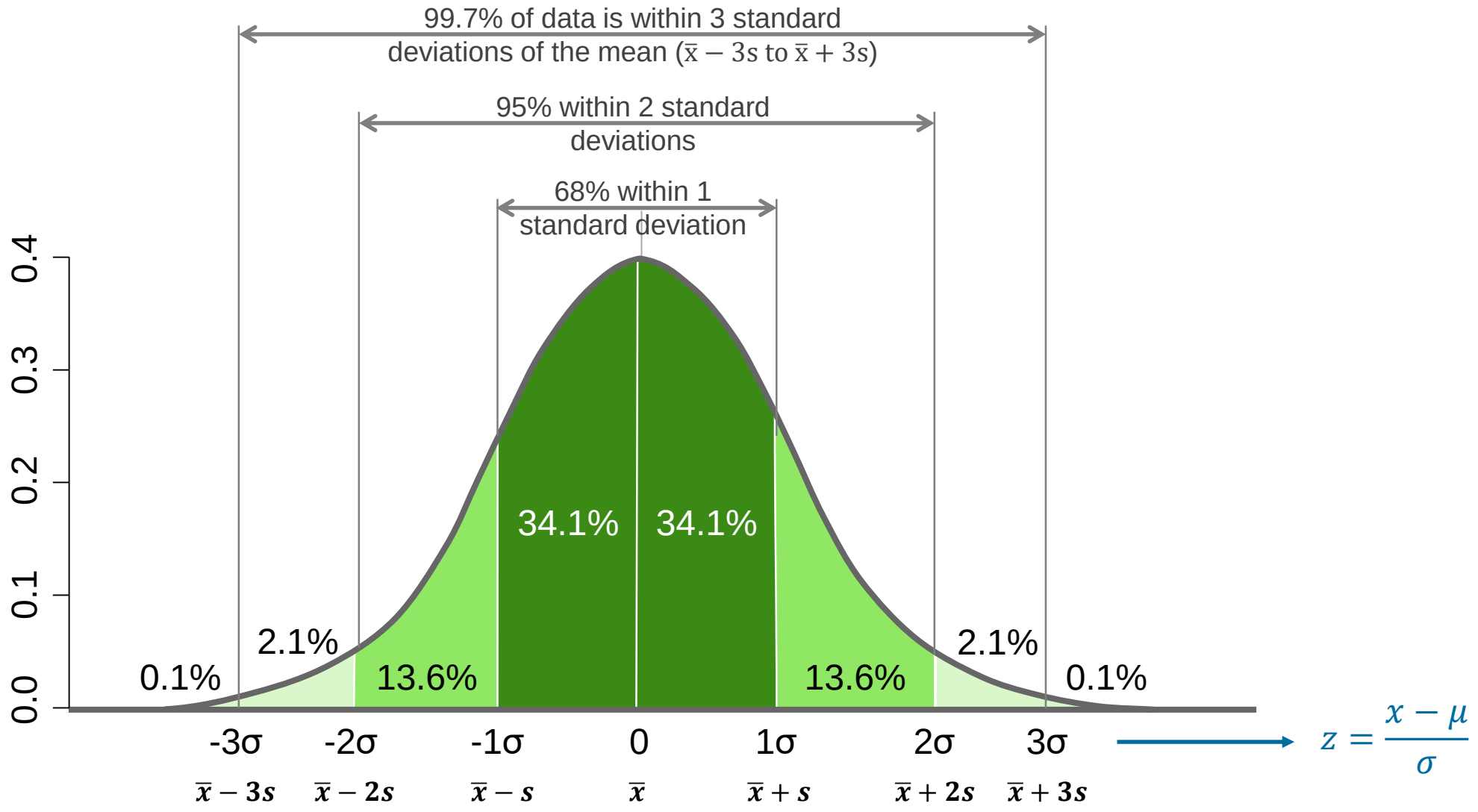


Summary of box plot features



Outliers

The empirical rule



Rules of thumb for detecting outliers using Z-scores

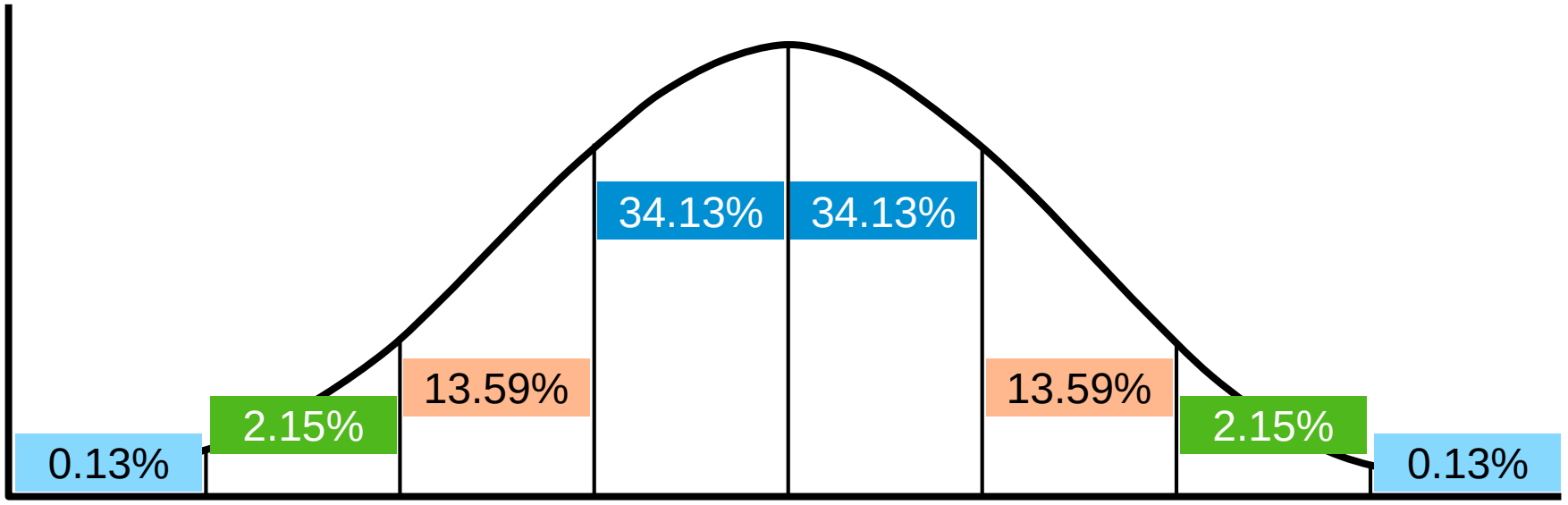
-1 SD = 55
-2 SD = 45
-3 SD = 35

65

+1 SD = 75
+2 SD = 85
+3 SD = 95

Possible Outliers
|z| > 2

Outliers
|z| > 3



Summary

- This lesson has introduced you to some simple but powerful methods to detect outliers.
- You have seen how the interquartile range, box plot, and empirical rule can be used to test for outliers.
- The empirical rule and box plot methods both establish rule-of-thumb limits outside of which a measurement is deemed to be an outlier.
- Usually, the two methods produce similar results. However, the presence of one or more outliers in a dataset can inflate the computed value of the standard deviation. Consequently, it will be less likely that an errant observation would have a Z-score larger than |3|. In contrast, the values of the quartiles used to calculate the intervals for a box plot are not affected by the presence of outliers.

Thank you.

Contact information:

open@sap.com

Follow all of SAP

www.sap.com/contactsap

THE BEST RUN SAP®