STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

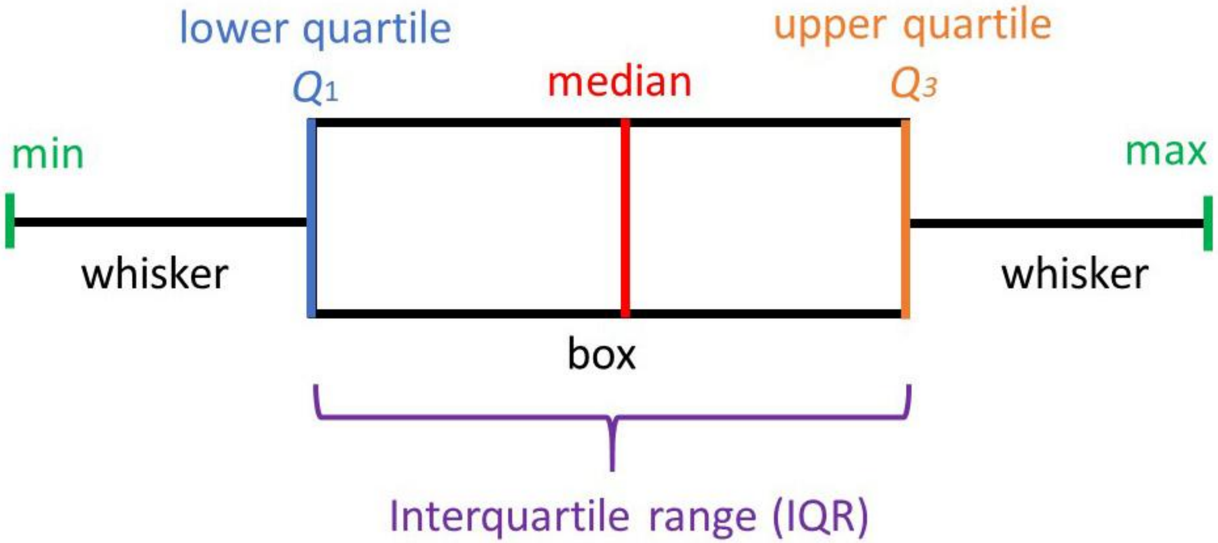
Data Visualization and Interpretation - Boxplot

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Boxplot

A **boxplot is a graphic that presents the median, the first and third quartiles, and any** outliers that are present in a sample.

It shows the distribution of a set of data along a number line, dividing the data into four parts using the median and quartiles.



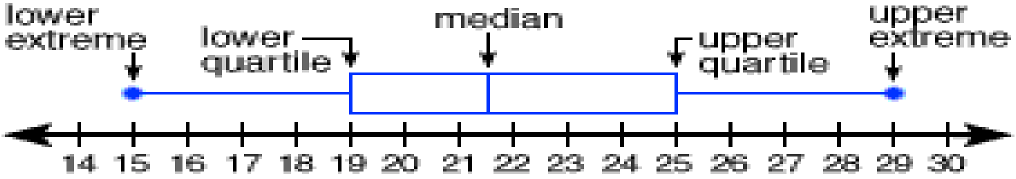
Source: Simply Psychology.com **Text Book – Chapter 1.3 – Pg. No - 33**

Why Boxplot?

A **box and whisker plot** is a way of summarizing a set of data measured on an **interval scale**.

It is a graph that presents information from a **five-number summary**.

It is often used in explanatory data analysis.



This type of graph is used to show the shape of the distribution, its central value, and its variability.

Box and whisker plots are ideal for comparing distributions because the centre, spread and overall range are immediately apparent.

Source: www.150.statcan.gc.ca Text Book – Chapter 1.3 – Pg. No - 33

Why Boxplot?

It **does not show a distribution** in as much detail as a stem and leaf plot or **histogram** does, but is especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set.

Box and whisker plots are also very useful when large numbers of observations are involved and when two or more data sets are being compared.

Shows how the values in the data are spread out.

Source: www.150.statcan.gc.ca

Boxplot

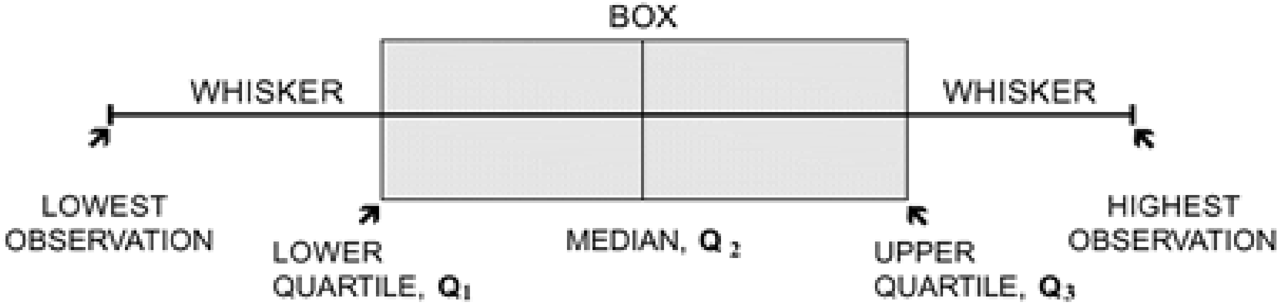
In a box and whisker plot:

- the **ends of the box** are the **upper and lower quartiles**, so the box spans the **interquartile range**.

- the **median** is marked by a **vertical line** inside the box.

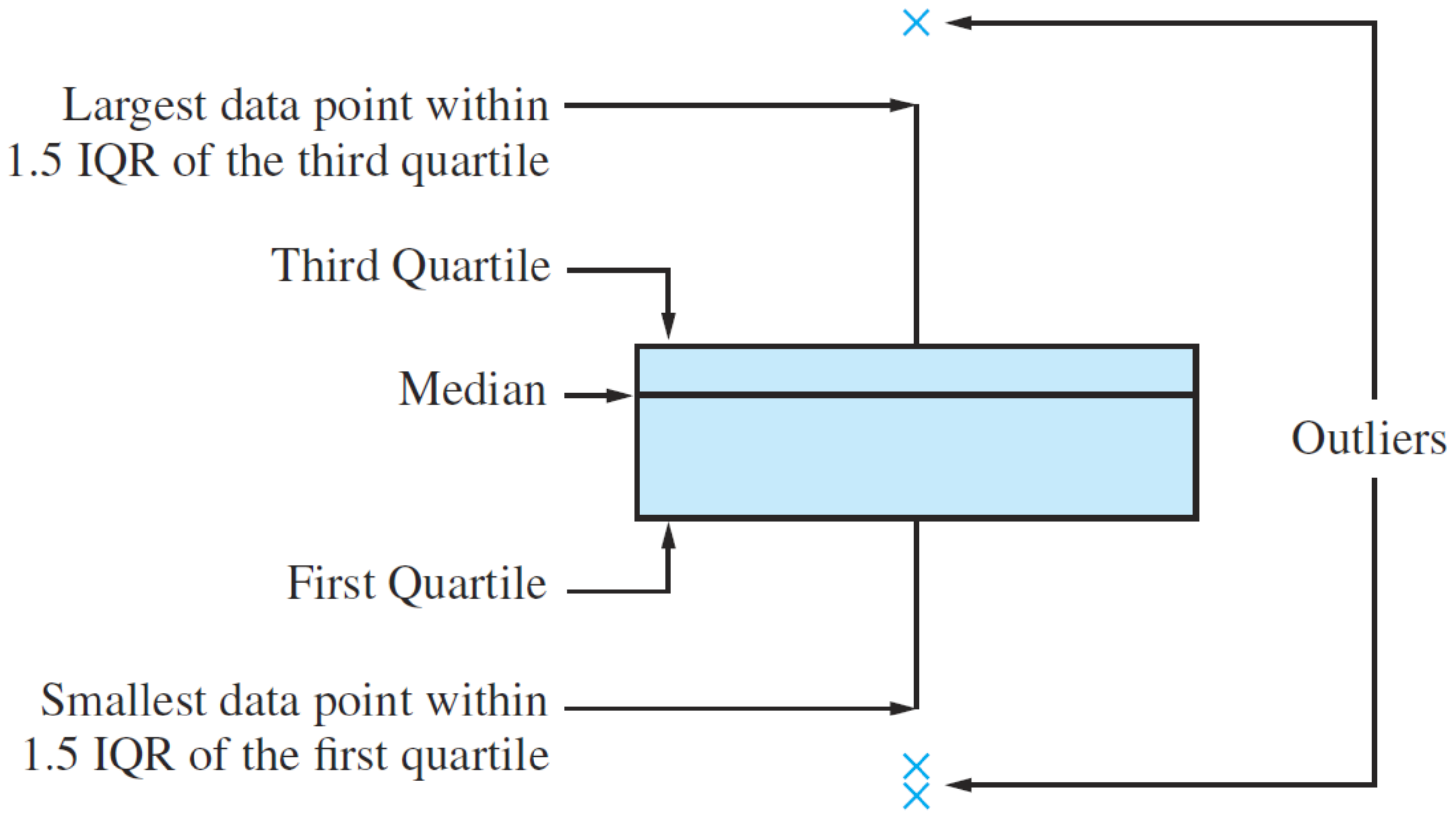
- the **whiskers** are the **two lines** outside the box that **extend to the highest and lowest observations.**

Figure 1. Box and whisker plot

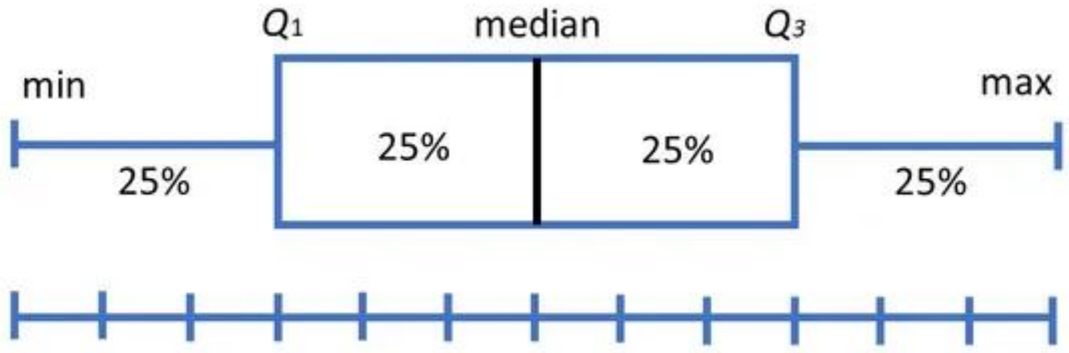


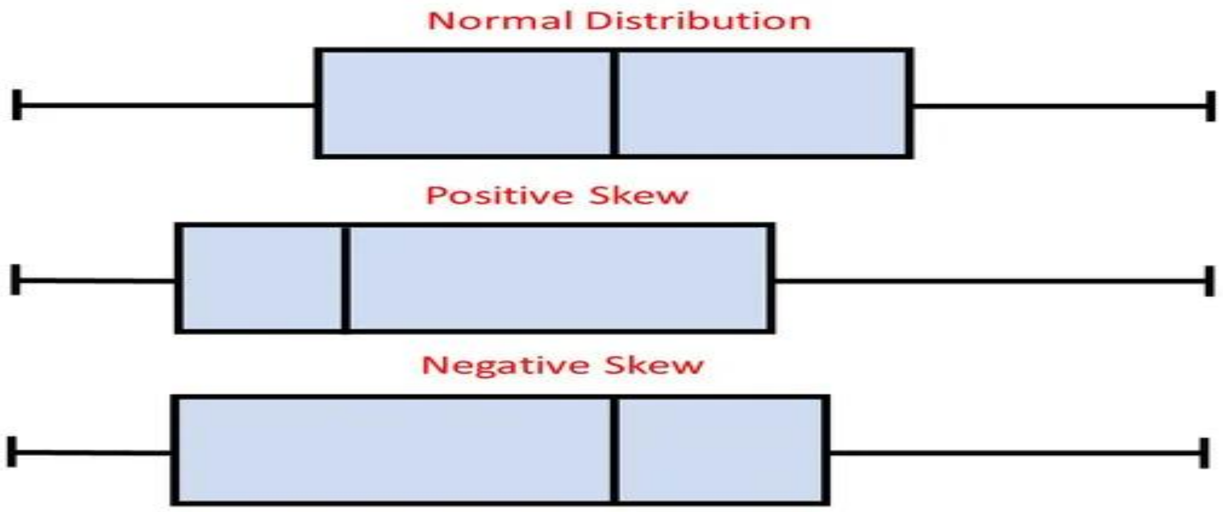
Source: www.150.statcan.gc.ca

Detailed Anatomy of Boxplot


Largest data point within
1.5 IQR of the third quartile
Third Quartile
Median
First Quartile
Smallest data point within
1.5 IQR of the first quartile
Outliers

Boxplot – Symmetry Vs Non-Symmetry





Source: https://www.simplypsychology.org/

Steps to Construct Boxplot

Step 1: Order the data from smallest to largest.

Step 2: Find the median.

Step 3: Find the quartiles.

Step 4: Complete the five-number summary by finding the min and the max.

Step 5: Making a boxplot.

a) Scale and label an axis that fits the five-number summary.

b) Make solid dots against Q1, Q2 and Q3 values above the number line.

c) Draw vertical lines from number line to those 3 points and complete the box .

d) Draw two horizontal lines from outside box(either side) to till the minimum and maximum value respectively. These lines are called whiskers.

Source: goLEANSIXSIGMA.com

Example

Consider the following example,

10	11	12	25	25	27	31	33
34	34	35	36	43	50	59	

Arrange the data in order

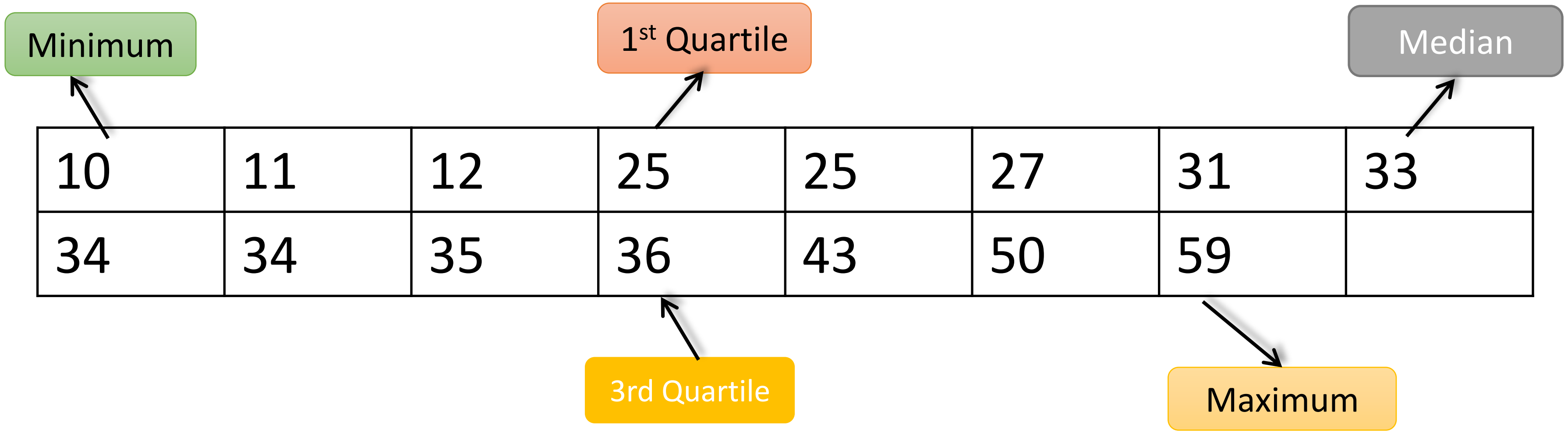
Position of the Median & Value= $\frac{n+1}{2}$ = 8 [33]

Position of the 1st Quartile & Value = 4 [25]

Position of the 3rd Quartile & Value = 12 [36]

Minimum Value = 10 Maximum Value= 59

Example – Five Number Summary



Minimum

1st Quartile

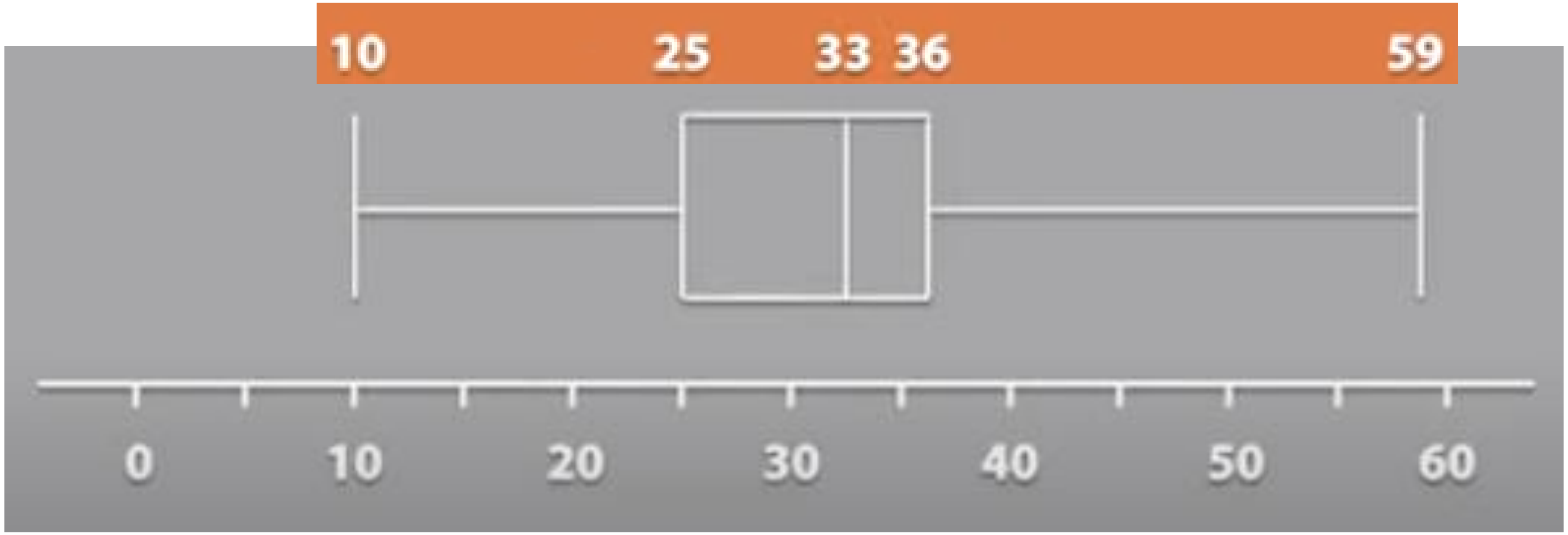
Median

10	11	12	25	25	27	31	33
34	34	35	36	43	50	59	

3rd Quartile

Maximum

Example – Boxplot Construction



Source: Simple Learning Pro

Example – Outlier Calculation

Outliers are points that are unusually large or small.

A data value is considered to be an outlier if,

DataValue $<$ $Q_1 - 1.5\ (IQR)$

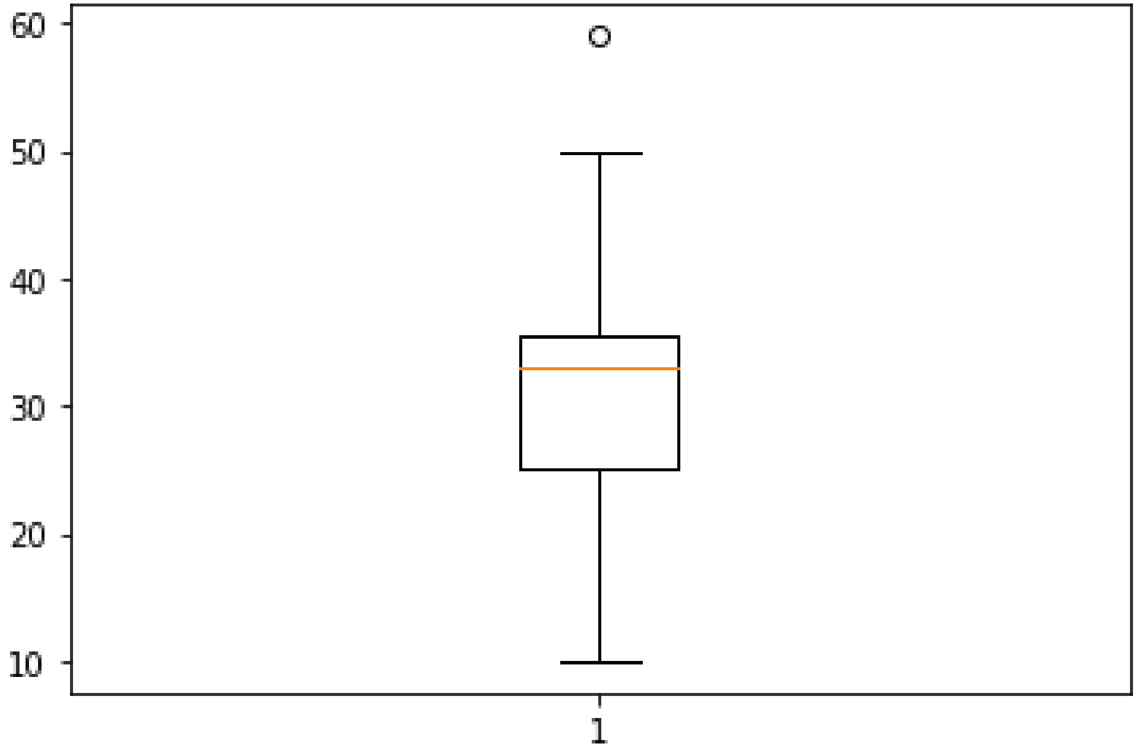
DataValue $>$ $Q_3 + 1.5\ (IQR)$

Let’s calculate the IQR $= Q_3 - Q_1 = 36 - 25 = 11$

Let’s substitute now,
$Q_1 - 1.5\ (11) = 8.5$

$Q_3 + 1.5\ (11) = 52.5$ (There will be one outlier as per the data)

Example – Resulting Boxplot with an Outlier

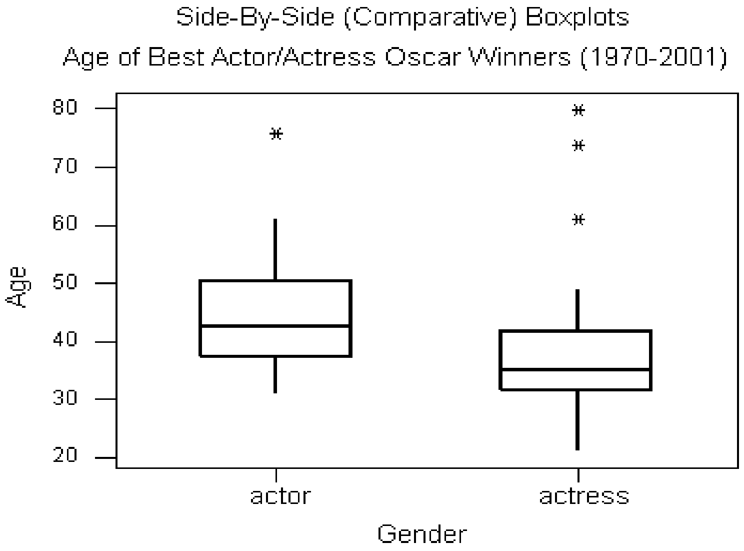


Comparative Boxplot

We can compare the boxplots of the two (or more) samples side-by-side.

This will allow us to compare how the medians differ between samples, as well as the first and third quartile.

It also tells us about the difference in spread between the two samples.



Summary

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.

Text Book – Chapter 1.3 – Pg. No - 34

THANK YOU

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Department of Computer Science and Engineering