Univariate Statistics

Class cases

Cars: Mileage & Gears & Weight &….

It is believed that mileage of a car is related to whether it has automatic transmission or not, examine this using mtcars data.

Using other information in the data analyse how other factors such as weight , engine type, horse power etc affect car mileage and each other

Clinical Trial

Analyse data from clinical trial of an Arthritis medicine [Data : Arthritis, Package: vcd], to find out whether the medicine is effective or not.

Find out whether gender or age of person has any effect on medicine performance

What is Univariate Statistics?

Summarising Behaviours of Data

- Univariate Statistics is study of individual variables as the name suggests

- Its not always possible to go through entire data, you'd rather like to look at few summary points which would give an idea/overview of the variables behaviour

Summary Statistics

Facets of the data

- To arrive at proper summary statistics we need to understand what facets of the data we need to look to have a complete overview:
  - Central Tendency
  - Variability
  - Shape

Central Tendency

- Central tendency means average or representative behaviour of the data

- Measures of central tendency are:
  - Mean
  - Median
  - Mode

Calculating Measures

- Mean $=\Sigma(X_1+X_2+X_3....)/n$

- Median : middle most value when all the values are sorted
  - In case of even number of values, average of two middle most values is taken as median

- Mode: Value which is most frequent. This is used for categorical variables mostly

Properties of the measures

- Mean is sensitive to extreme values, Median isn't.
- Mean, Median ; Both take unique values only
- Mode can take multiple values

Variability

- Equal central tendency doesn't mean similar all round behavior

- Data values can be different in terms of spread around the central tendency

Measures of Variability

- Range
- Standard Deviation / Variance
- Mean Absolute Deviation (MAD)
- Inter Quartile Range (IQR)
  - Q2 (second quartile aka median) divides data into two parts. Q1 :first quartile divided first part into two equal parts and Q3: third quartile does the same for second part.

Properties of Measures

- Range is most sensitive to extreme values

- Std , Variance and MAD are sensitive to extreme values as well

Properties of Measures

- IQR ignores leading and trailing values, hence is not sensitive to extreme values

- MAD is as good a measure of variability as variance , only issue is that its difficult to manipulate algebraically

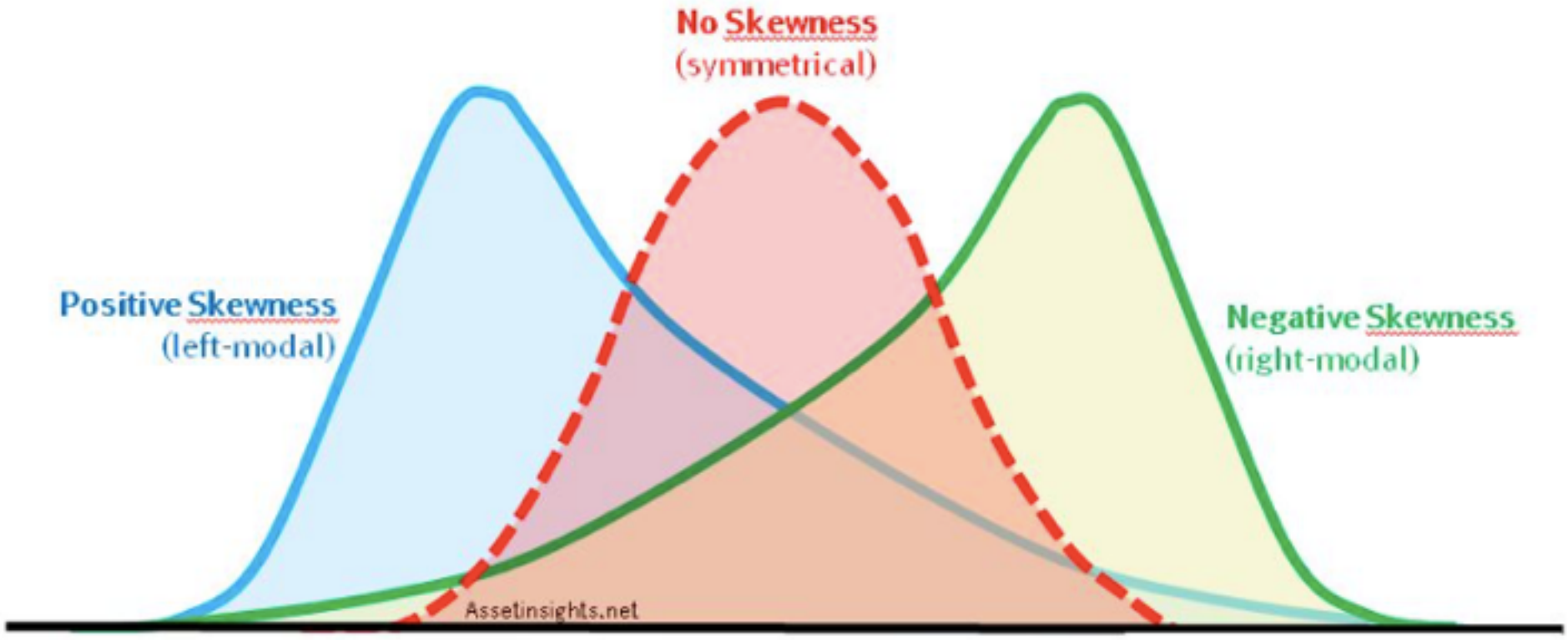
Shape

- Central tendency and variability summary stats give you an idea about how your data is centered and spread but it doesn't tell you how frequent particular values are in your data

Types of shapes & Numerical Measure

- Symmetric
- Positively Skewed
- Negatively Skewed
- Measure:
  - Skewness

Types of shapes

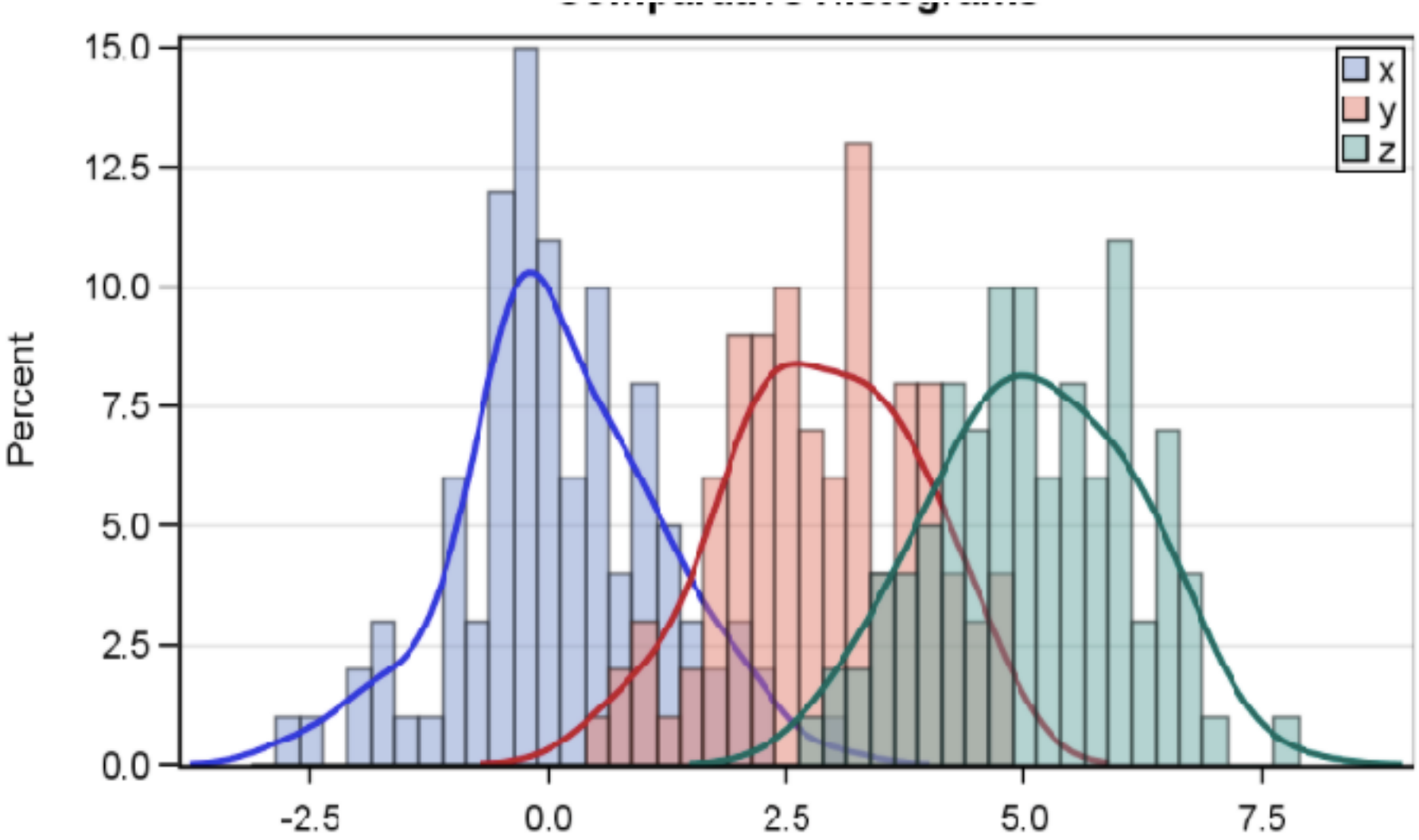


Visualising Shapes

- Histograms: These are simple bar charts indicating frequency of data values in your data

- Box Plot: This shows relative positions of quartiles in your data, thus giving you an idea about the shape of the data

Histogram

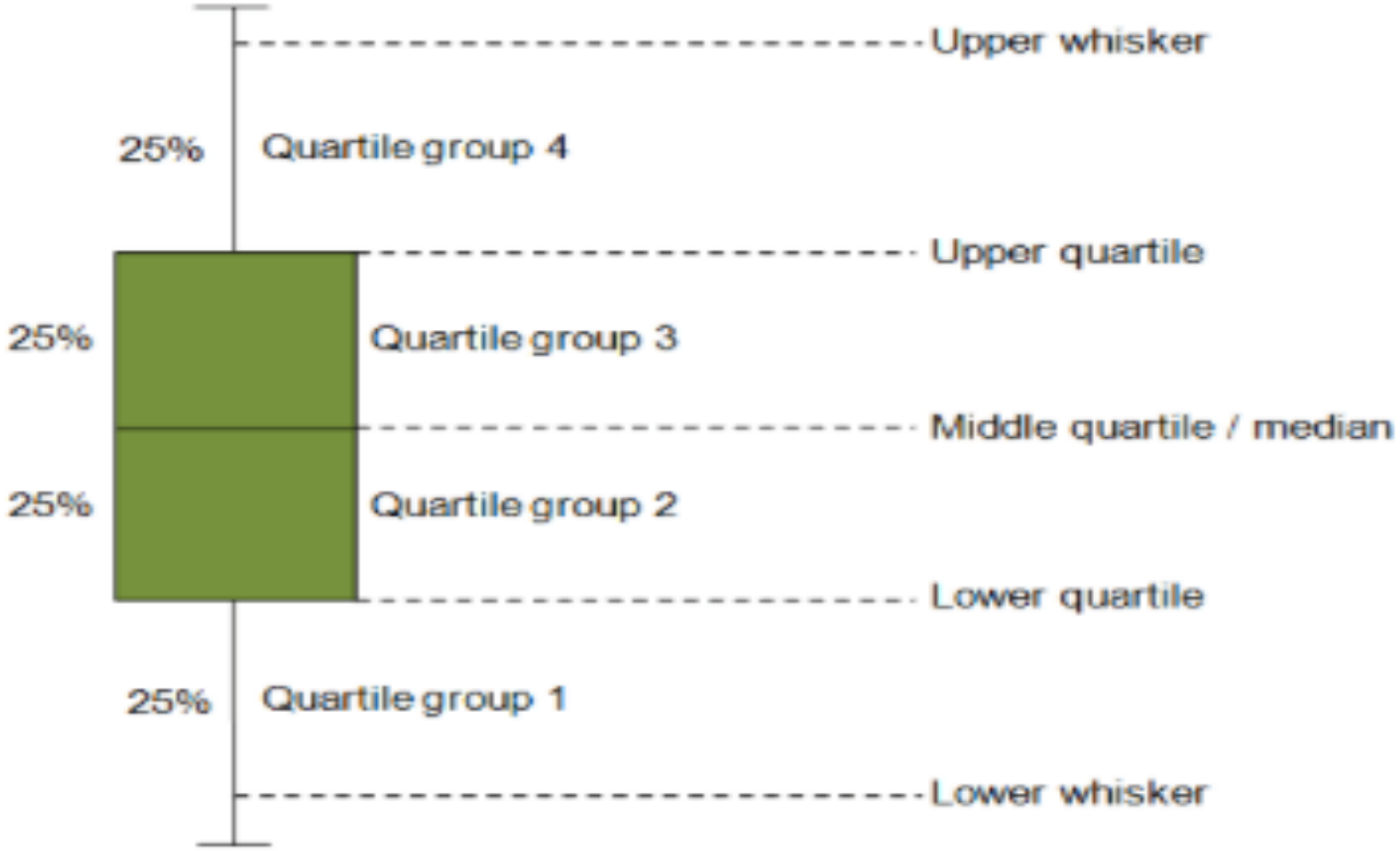


About Histograms

- Continuous numeric vars have too many unique values to make a proper histogram [bar chart with frequencies], for a better visual, we club neighbouring values in broader classes .

- These classes can be made finer when we have large amount of data

Boxplot



Outliers

- They are defined for numeric variables

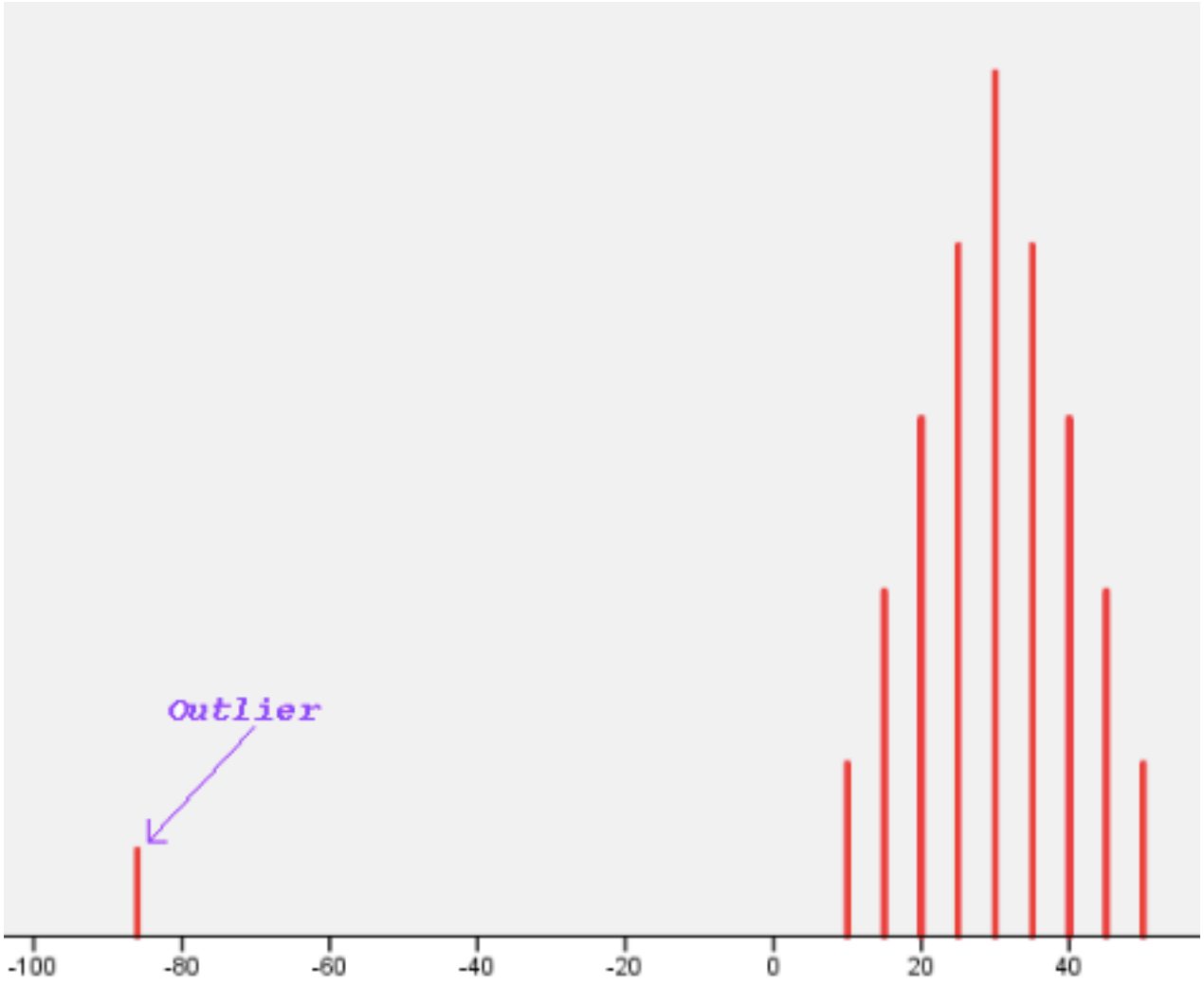
- Values in the data which are too different from rest of the data, meaning they are either too small or too large

Quantifying Outliers

- We can quantify these “too small” and “too large” values by defining an acceptable range of values for data

- Standard example of ranges:
  - Mean ± N * Std
  - $[Q_1 - N*IQR, Q_3 + N*IQR]$

Outliers



Implementation in R

Implementation in R