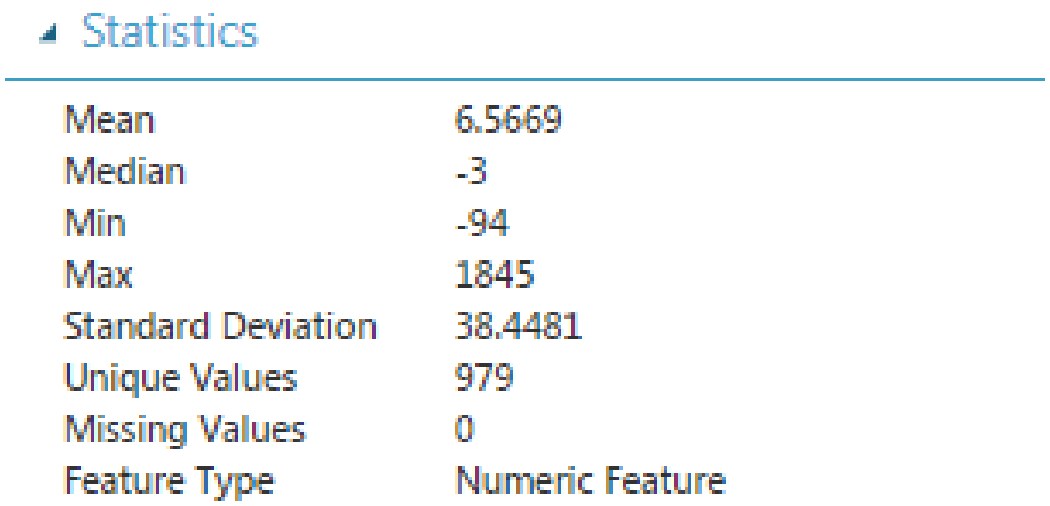
Module 1 :
Descriptive Statistics

Topics

- Defining Data
- Histogram and Skewness
- Descriptive Statistics with Analysis ToolPak
- Boxplots
- Categorical Data, PivotTables, and PivotCharts
- Summarizing Hierarchical Data
- 80-20 Rule and Pareto Charts

Descriptive statistics

- Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or sample population.
- Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).



Statistics

Mean	6.5669
Median	-3
Min	-94
Max	1845
Standard Deviation	38.4481
Unique Values	979
Missing Values	0
Feature Type	Numeric Feature

Defining Data

- Individuals and Variables

Individuals are the objects described by the data set.

Individuals might be people, stocks, cities, etc.

A variable is a characteristic of an individual

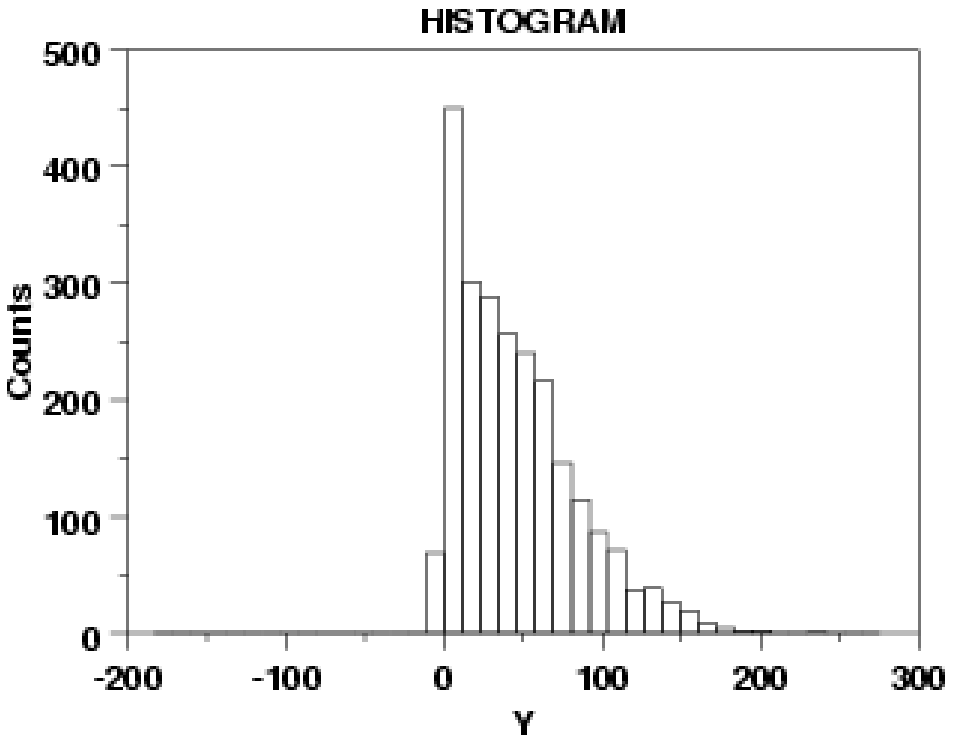
- Numerical and Categorical Data
- Categorical: Nominal - Ordinal
- Numerical: Discrete vs Continuous

Histogram and Skewness

- Used to to graphically summarize the distribution of a univariate data set
- splitting the range of the data into equal-sized bins (called classes)
- Information can be obtained from histogram:
  - center (i.e., the location) of the data;
  - spread (i.e., the scale) of the data;
  - skewness of the data;
  - presence of outliers; and
  - presence of multiple modes in the data.

Histogram and Skewness

- Histogram Skewness:
  - Normal
  - Symmetric with Short or Long Tailed
  - Skewed (Non-Symmetric) Right or Left;
  - Symmetric with Outlier

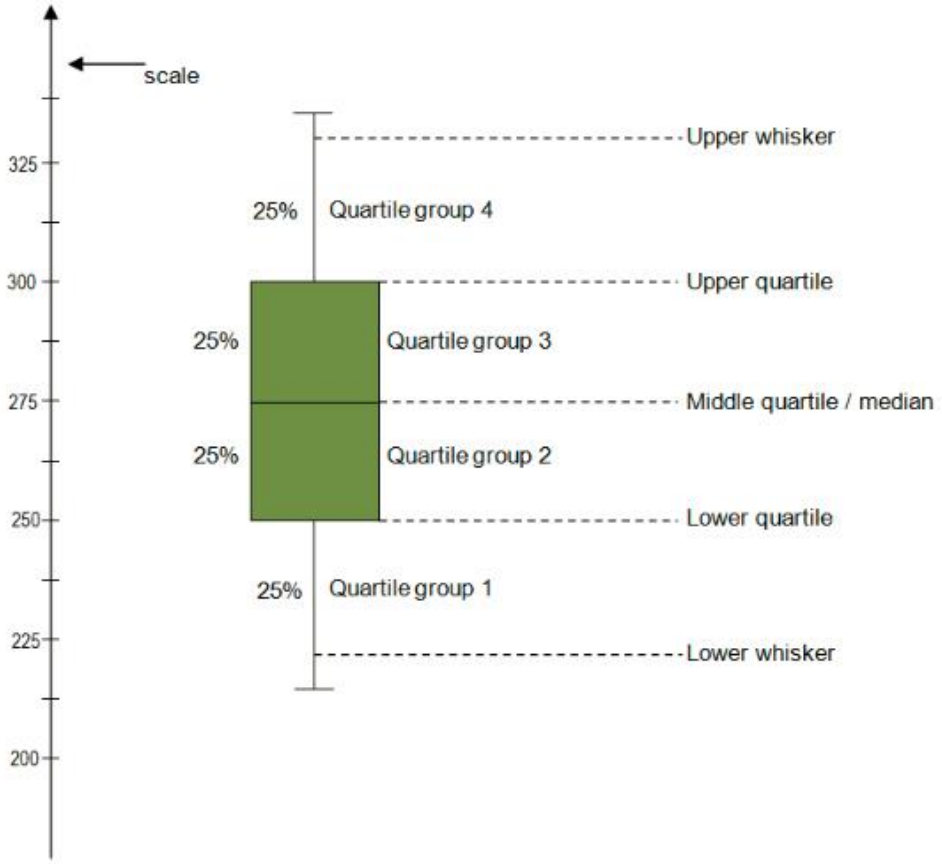


Boxplots

Interpreting box plots/Box plots in general

- Box plots are used to show overall patterns of response for a group. They provide a useful way to visualise the range and other characteristics of responses for a large group.
- They enable us to study the distributional characteristics of a group of scores as well as the level of the scores.

Boxplots - example

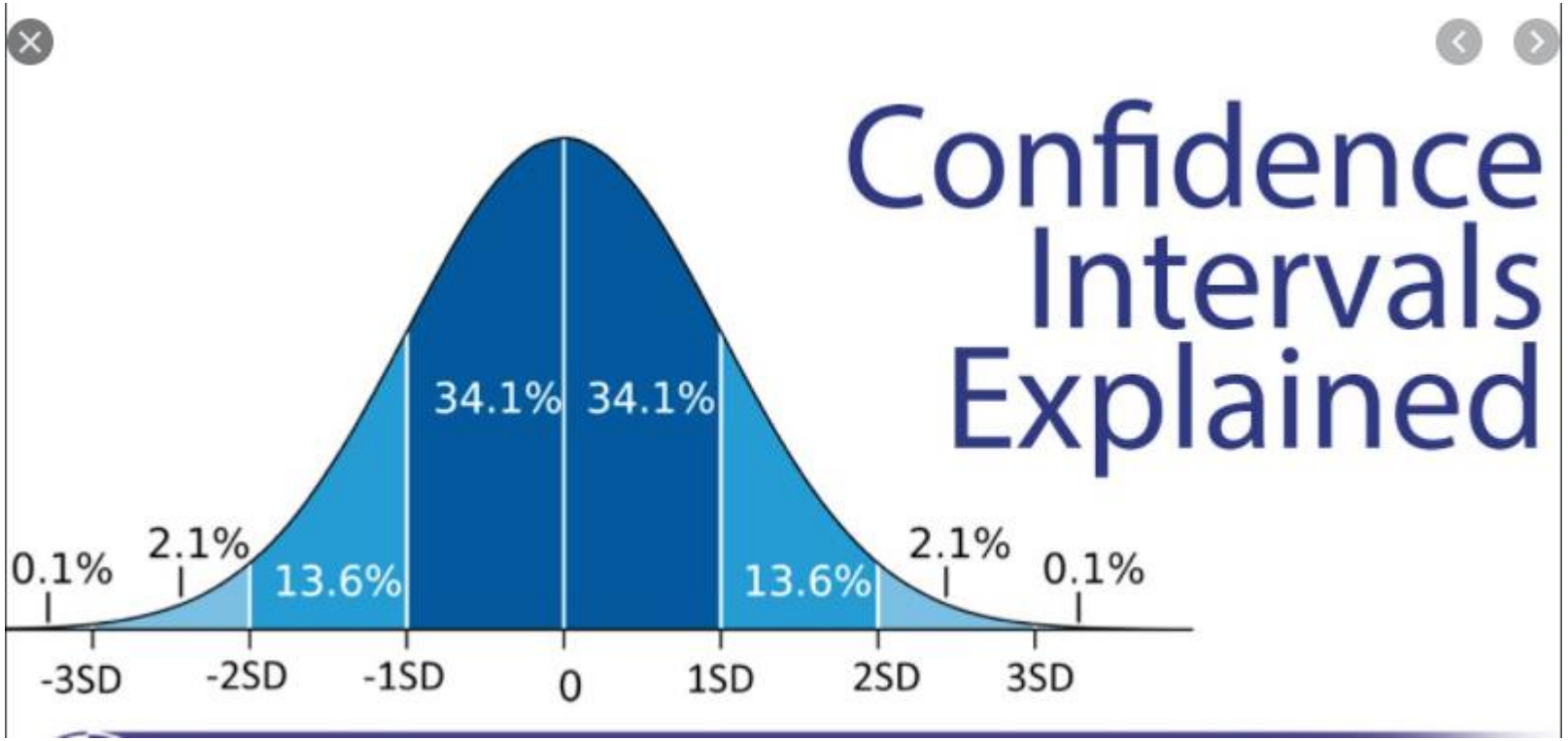


- Data are sorted by value and divided into 4 Quartiles, each quartile contains 25%
- Quartile2 & Quartile3 (50%) are presented in the box

Boxplots - Definitions

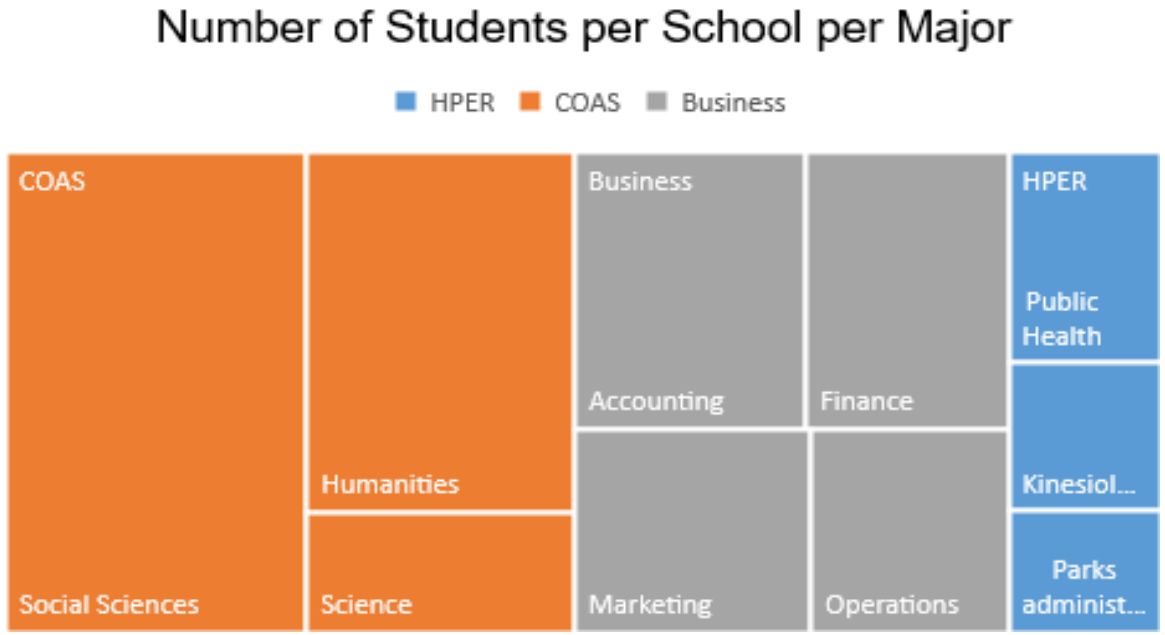
- **Median (middle quartile)**
  Marks the mid-point of the data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less.
- **Inter-quartile range**
  The range of scores from lower to upper quartile is referred to as the inter-quartile range. The middle 50% of scores fall within the inter-quartile range.
- **Upper quartile**
  Seventy-five percent of the scores fall below the upper quartile.
- **Lower quartile**
  Twenty-five percent of scores fall below the lower quartile.
- **Whiskers**
  The upper and lower whiskers represent scores outside the middle 50%. Whiskers often (but not always) stretch over a wider range of scores than the middle quartile groups.

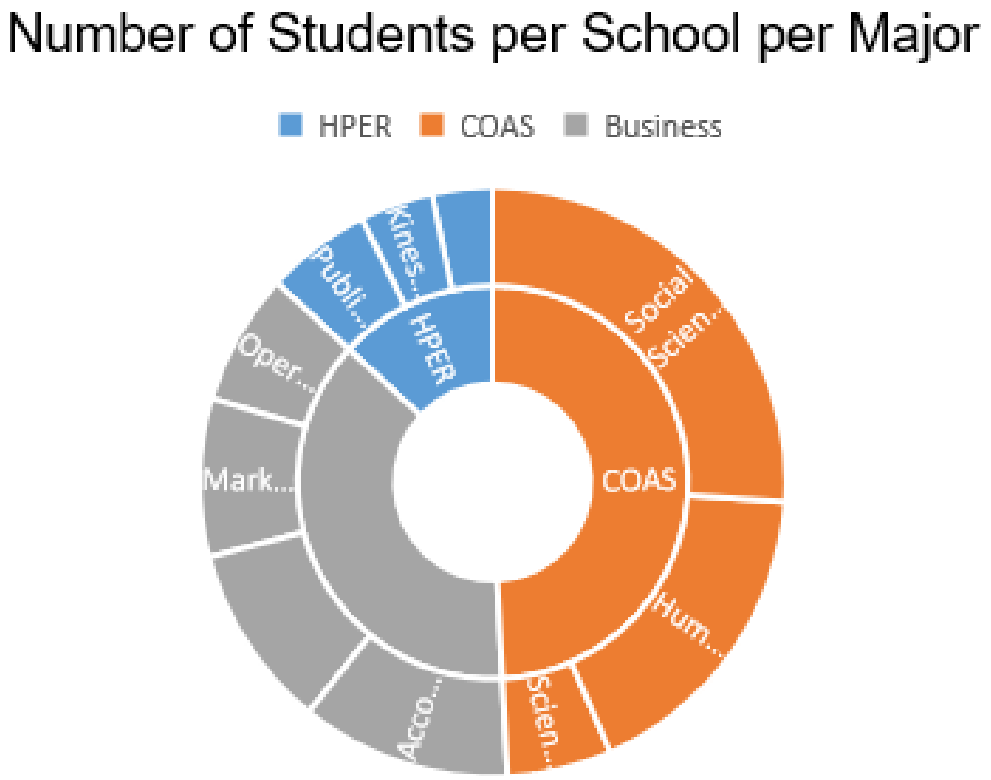
Confidence Intervals



Summarizing Hierarchical Data

- Summarizing Hierarchical Data with Treemap & Sunburst Charts





80-20 Rule and Pareto Charts

- Known as the Pareto principle
- The 80/20 Rule means that in any situation 20% of the inputs or activities are responsible for 80% of the outcomes or results.
- 80% of revenues are generated by 20% of customers.
- 80% of our quality issues occur with 20% of products.
- 20% of employee are responsible for 80% of sick days.

Pareto Charts

- A Pareto chart, named after Vilfredo Pareto
- Is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line