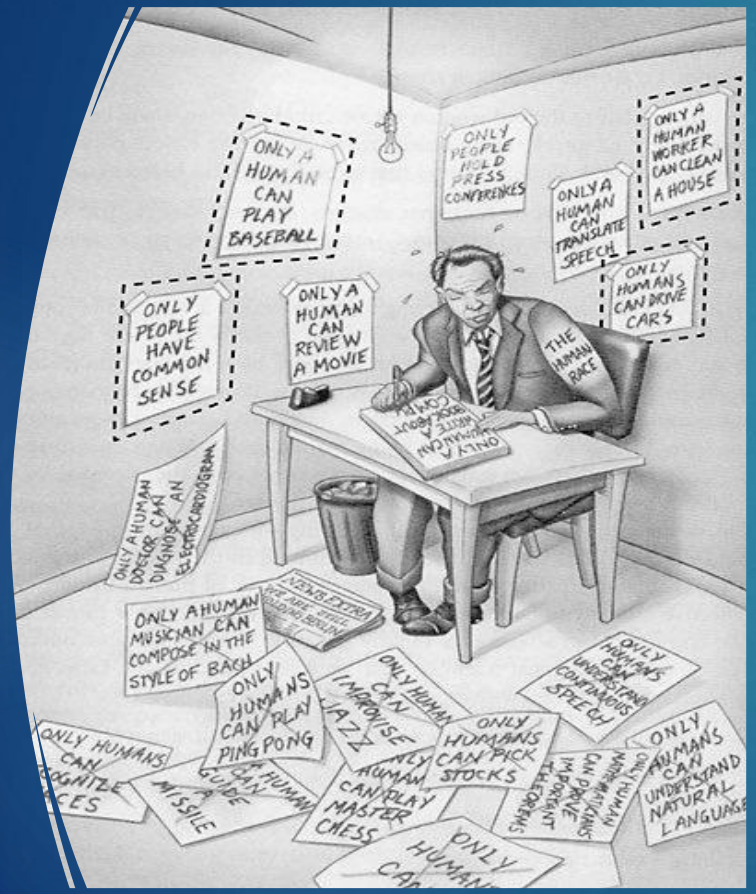
Certificate Program in Machine Learning & Artificial Intelligence: Batch-3

NEENA PANDEY

IIM VISAKHAPATNAM

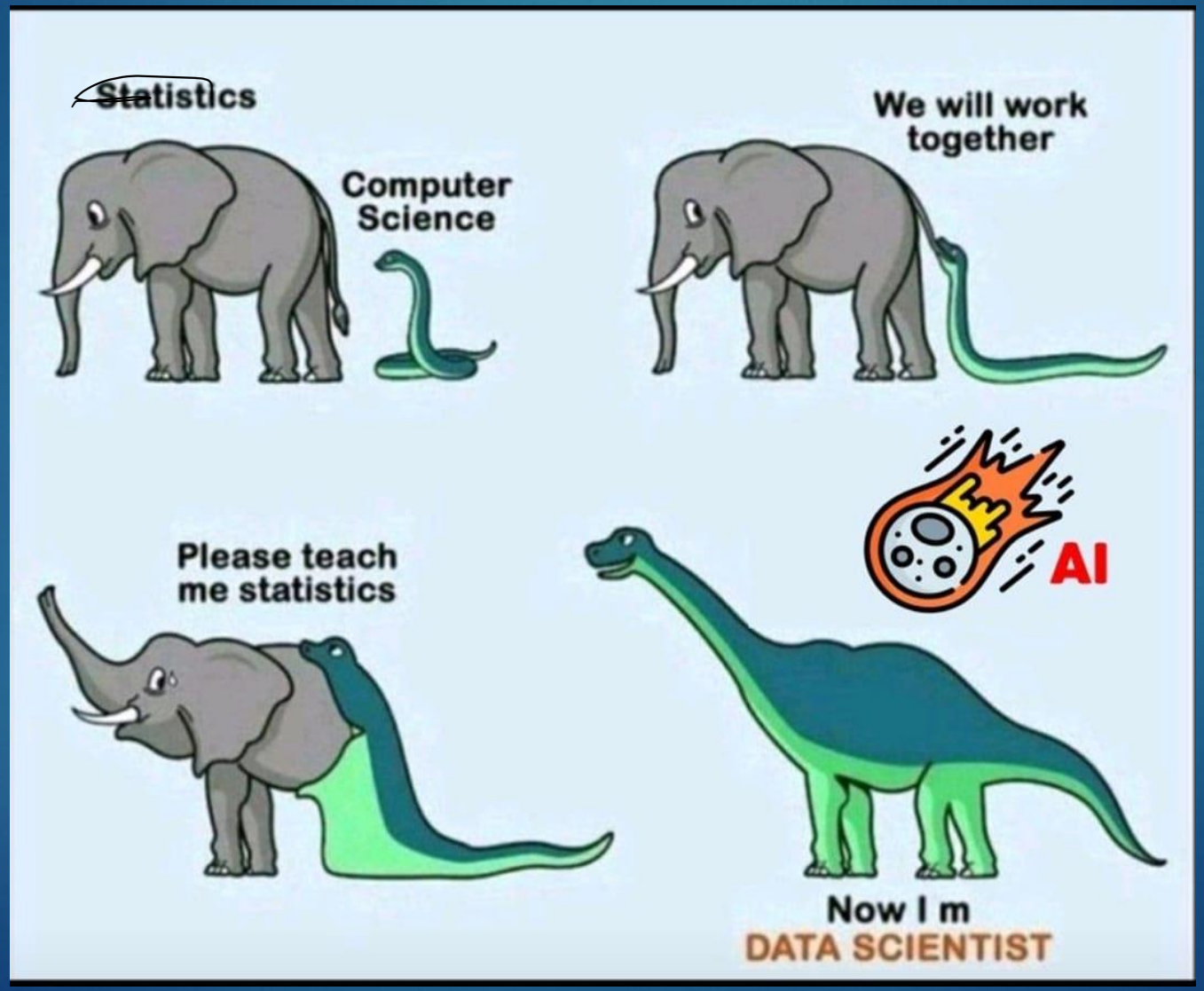
AI & ML



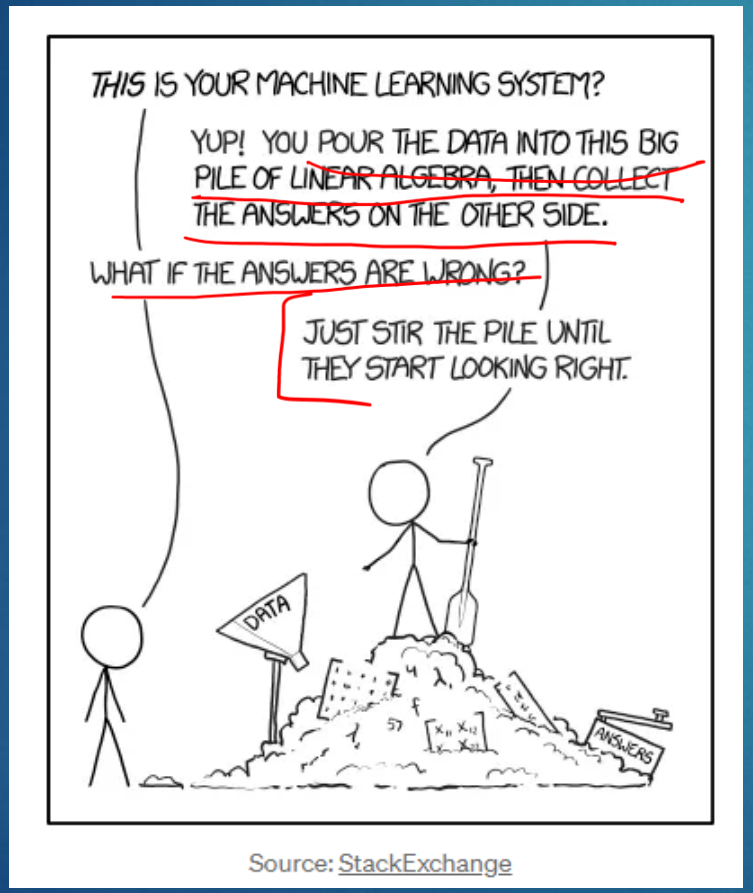
The 50 most attractive nationalities revealed: India is No.1, USA comes second and Britain has the most handsome men (while AI images show 'beautiful people' in each country)

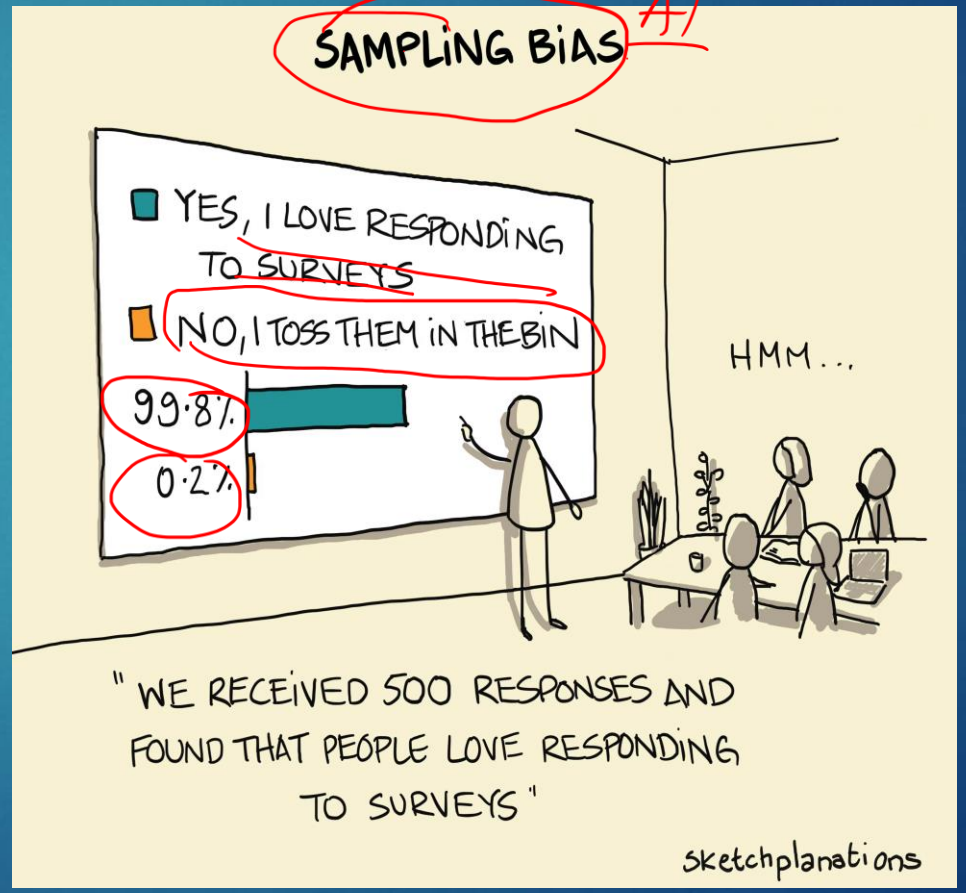
By Ted Thornhill, Mailonline Travel Editor
12:42 03 Mar 2023, updated 12:13 04 Mar 2023

AI/ML, Analytics & Statistics



AI/ML & Statistics





Module 1: A Primer on Statistical Concepts

Over the course of Basic Statistics sessions

- Descriptive Statistics
  - Types of measurement, Tabular & Pictorial representation
  - Numerical measures – Central tendency (Location) and Dispersion
- Probability
  - Probability (Basic & Conditional), Bayes' Theorem
  - Probability distributions
- Hypothesis Testing
  - Point & Interval Estimation
  - Hypothesis Testing

Statistics

- Refers to *numerical facts* such as averages, medians, percentages, and maximums that help us understand various business and economic situations.
- Also refers to the *art and science* of collecting, analyzing, presenting, and interpreting data.
- Business Applications
  - Statistical sampling procedures used by public auditing firms
  - Financial advisors may use PE ratio and dividend yields to guide their investment advice
  - Information Systems: To monitor performance and anomaly in computer networks

Data

- Scales of measurement
  - Nominal, Ordinal, Interval, and Ratio
- Data Classification
  - Categorical & Quantitative Data
  - Cross-sectional vs. Time Series Data
  - Discrete and Continuous Data
- Data Sources
  - Internal company records – across functional departments
  - Govt. agencies
  - Business Database services
  - Industry associations, Internet etc.

Types of Statistical Studies

- Observational
  - No attempt is made to control or influence the variables of interest
  - Example: Survey
- Experimental
  - Identification of the variable of interest; Performed under controlled conditions
  - Other variables are identified and controlled or changed to study their impact on the variable of interest
  - Example: Agriculture, Medical experiments, A/B testing for websites or other digital products

Descriptive Statistics: Tabular & Graphical Representation

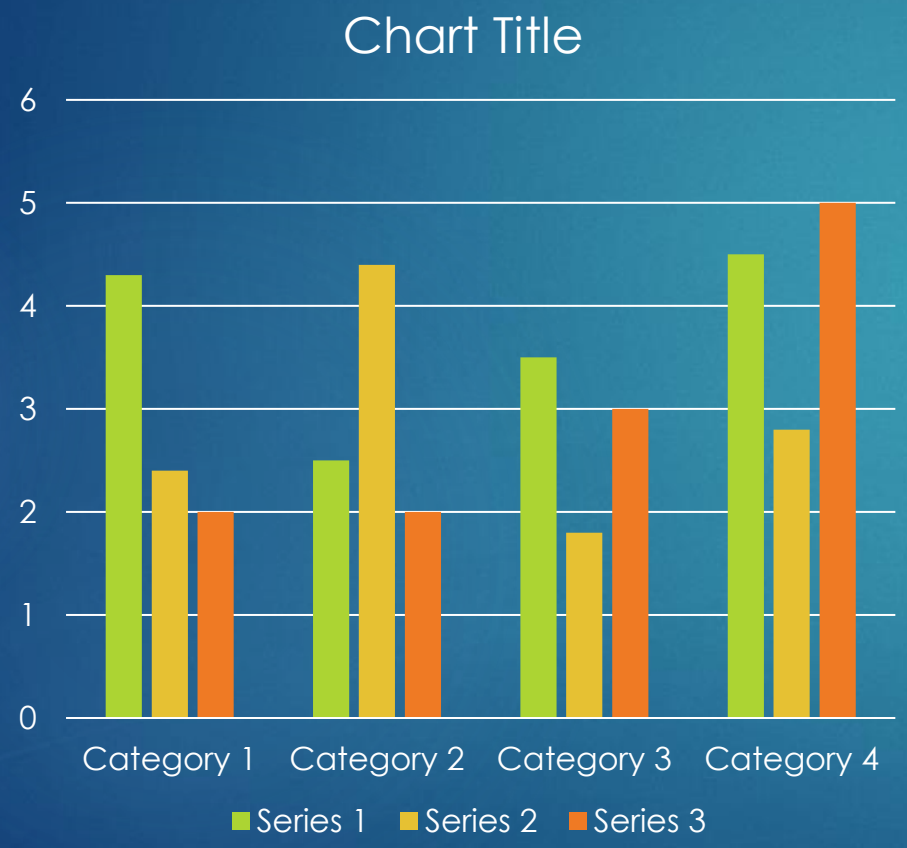
Categorical Data Summarization: Graphs & Tables

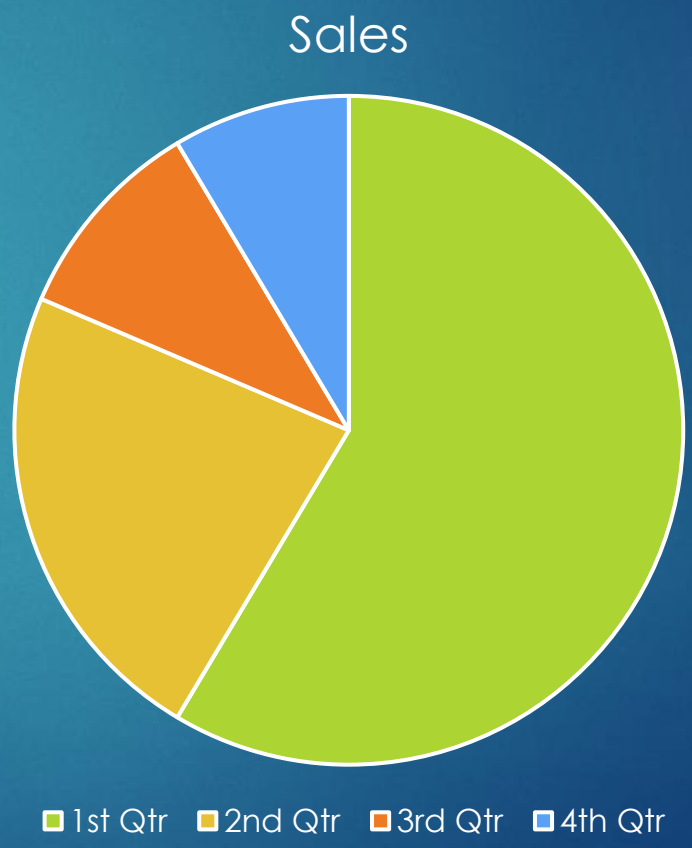
- Frequency Distribution (SoftDrink.xlsx, Ratings.xlsx)
  - Tabular summary of data showing the number (frequency) of observations in each of several *non-overlapping categories* or classes.
- Relative Frequency distribution/ Percentage frequency distribution
- Bar Chart
  - A graphical display for depicting qualitative data
  - Using a bar of fixed width drawn above each class label
- Pie Chart
  - Relative frequency and percent frequency distributions for categorical data.

Frequency Distribution: Tabular

Ratings	Frequency	Relative Frequency	Percent Frequency
Poor	2	0.10	10%
Below Average	3	0.15	15%
Average	5	0.25	25%
Above Average	9	0.45	45%
Excellent	1	0.05	5%
Total	20	1.00	100%

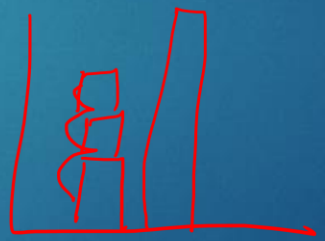
Bar Chart & Pie Charts





Quantitative Data Summarization: Graphs & Tables

- Only one variable of interest
  - Frequency, Relative Frequency, and Percent Frequency Distributions
  - Determine the number of nonoverlapping classes, width of class, class limits
  - Histogram
    - Horizontal axis – variable of interest; Vertical axis – frequency, relative frequency, percent frequency
  - Cumulative Distributions – Facebook adoption over the years
- Relationship between two variables
  - For comparisons between the two:
    - Side-by-side Bar Chart – for comparing two variables
    - Stacked Bar Chart – to compare relative frequency
  - For relationships between the two:
    - Scatter plot
    - Trendline

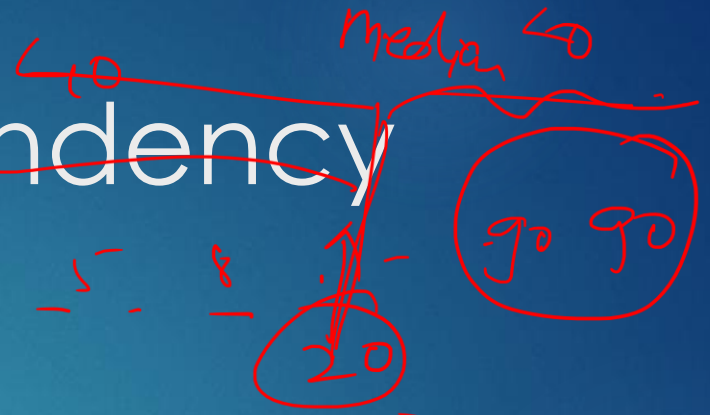


Descriptive Statistics: Numerical Measures

Quantitative Data Summary: Numerical Measures

- Measures of Central Tendency (Location)
- Measures of Variability (Dispersion)
- Detecting Outliers
- Measures of Association

Measures of Central Tendency



- Mean, Median, Mode, Weighted Mean, Percentiles, Quartiles
- Mean
  - Sample mean $\bar{x}$ is the point estimator of the population mean, μ.
- Median
  - Value in the middle of a data set when the data items are arranged in ascending order.
  - When is it preferred? – When a data set has extreme value
- Mode
  - Value that occurs with the greatest frequency

$$\bar{x} = \frac{\sum x_i}{n}$$

Measures of Central Tendency

- Weighted Mean
  - Mean where each observation has a weight that reflects its reflective importance
  - Choice of weights depends upon the application – for e.g., number of credits earned for each grade

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

where: x_i = value of observation i

w_i = weight for observation i

  - Another example?
  - Spot-market supplier or regular supplier ratio

Measures of Central Tendency

- Percentiles
  - Provides information about how the data are spread over the interval from the smallest value to the largest value.
  - p^{th} percentile of a data set is a value such that at least p percent of the items take on this value or less and at least (100 – p) percent of the items take on this value or more.
- Arrange the data in ascending order.
- Compute L_p, the location of the p^{th} percentile.

$$L_p = \left(\frac{p}{100}\right)(n+1)$$

Measures of Central Tendency

- 80th percentile for 70 observations

$$L_p = \left(\frac{p}{100}\right)(n+1) = \left(\frac{80}{100}\right)(70+1) = 56.8$$

- For e.g., NEET Score; Percentile = 635 + 0.8(649 – 635) = 646.2

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Measures of Central Tendency: Quartiles

- First Quartile (Q_1) = 25^{th} Percentile
- Second Quartile (Q_2) = 50^{th} Percentile
- Third Quartile (Q_3) = 75^{th} Percentile

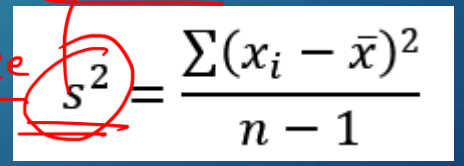
- Second Quartile is same as ??

Measures of Variability (Dispersion)

- Think of a business case where even if you know measure of location, measure of variability add significantly high value!
- Examples
  - Suppliers' average delivery time and variability in delivery time
  - Average return on investment and variation in investment return
- Common measures
  - Range, Interquartile Range, Standard Deviation, Coefficient of variation

Measures of Variability (Dispersion)

- Range
  - Difference between the largest and smallest data value.
  - Very sensitive to two specific data points
- Interquartile Range
  - Fixes earlier problem
  - Range for the middle 50% of the data – the difference between the third and first quartile
  - IQR = Q_3-Q_1 – Try with the given data
- Variance/Standard Deviation (Sq. root of variance)
  - Measure of variability that utilizes all the data
  - average of the squared deviations between each data value and the mean



Measures of Variability (Dispersion)

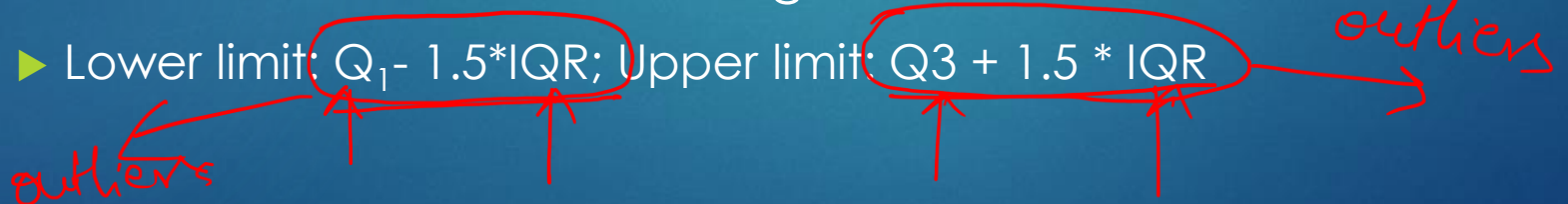
- Standard Deviation
  - More easily interpretable
- Coefficient of Variation
  - How large the deviation is in relation to the mean
  - What's a good value for Coefficient of Variation
  - When can this be useful?

$$\left[\frac{s}{\bar{x}} \text{ x } 100\right] \%$$

Detecting Outliers

Detecting Outliers – Why & How to detect

- Why
  - unusually small or unusually large value in a data set.
  - May be an incorrectly recorded value, incorrect inclusion, or correct value but an outlier data point
- Detection:
  - An observation in this range: More than 3* std_dev away
  - An observation outside this range:
    - Lower limit: Q_1 - 1.5*IQR; Upper limit: Q3 + 1.5 * IQR



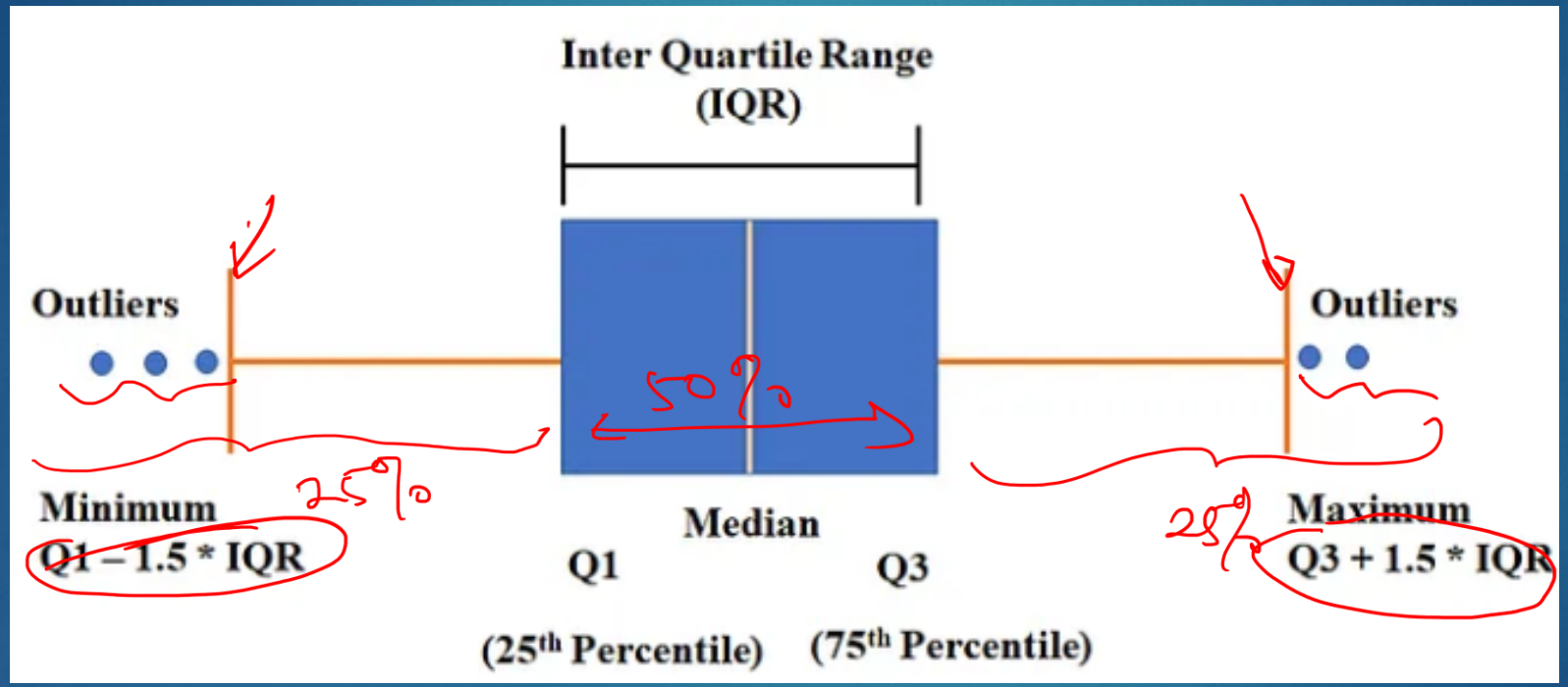
Five-Number Summaries & Box Plots

- Two tools that accomplish this are five-number summaries and box plots.
  - Smallest Value
  - First Quartile
  - Median
  - Third Quartile
  - Largest Value

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

- Box Plots

Box Plot



https://medium.com/@akashmishra77/box-plots-detect-and-remove-outliers-from-distribution-a124ee88cf3e

Measure of Association

- Till now, summarized data for one variable at a time
- What's the relationship between two variables?
  - Covariance & Correlation Coefficient
- Covariance
  - Measure of linear association between two variables

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Correlation Coefficient:

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

$-1 \quad \cdots \quad +1$

$+1 \rightarrow$ direction