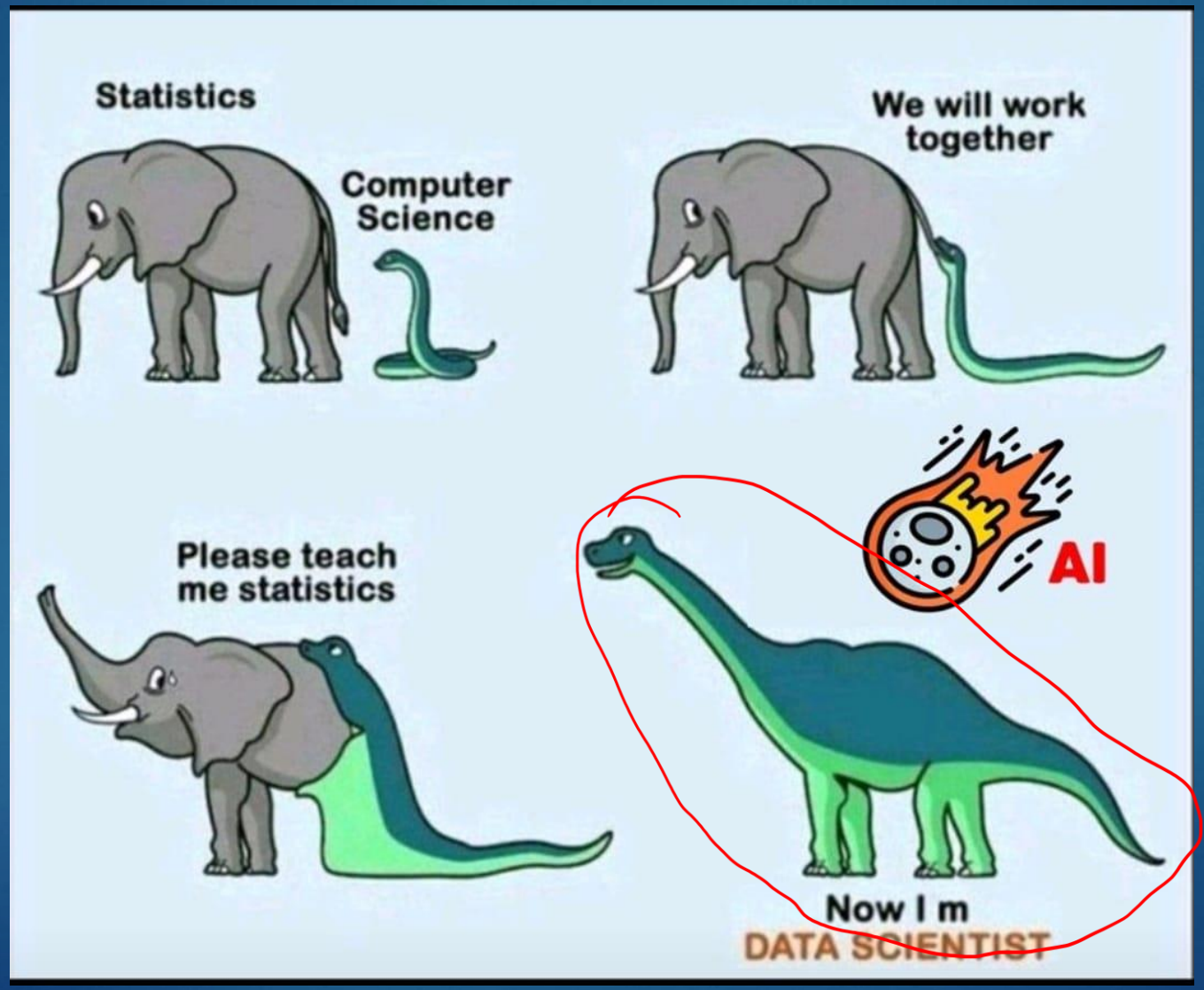
Module 1: Developing Managerial Mindset

POSTGRADUATE CERTIFICATE PROGRAM IN DIGITAL TRANSFORMATION STRATEGY & LEADERSHIP

NEENA PANDEY, IIM VISAKHAPATNAM

Module 1: S1: A Primer on Statistical Concepts

AI/ML, Analytics & Statistics



Over the course of 3 sessions

- Measures of Central Tendency, Dispersion & Association
- Probability (Basic & Conditional), Bayes' Theorem
- Probability distributions – Discrete & Continuous
  - Uniform, Normal

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

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.

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

- 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

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

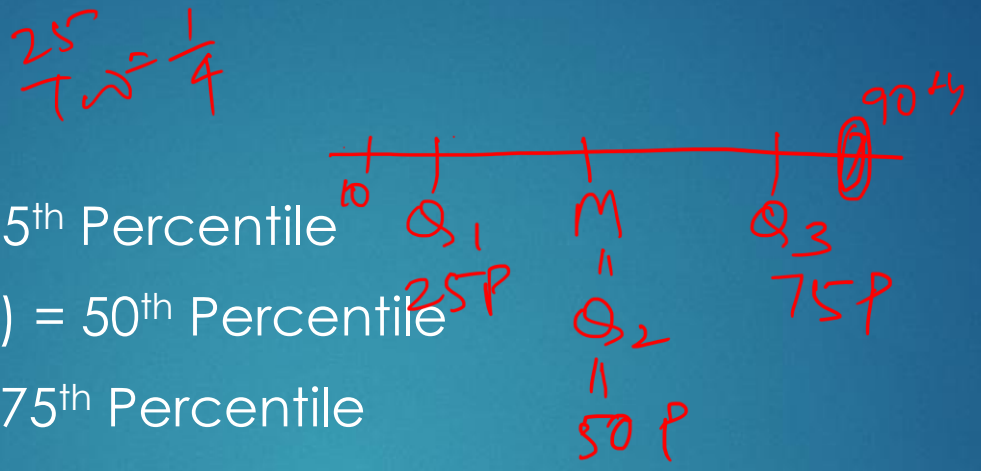
- 80th percentile for 70 observations

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

- 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

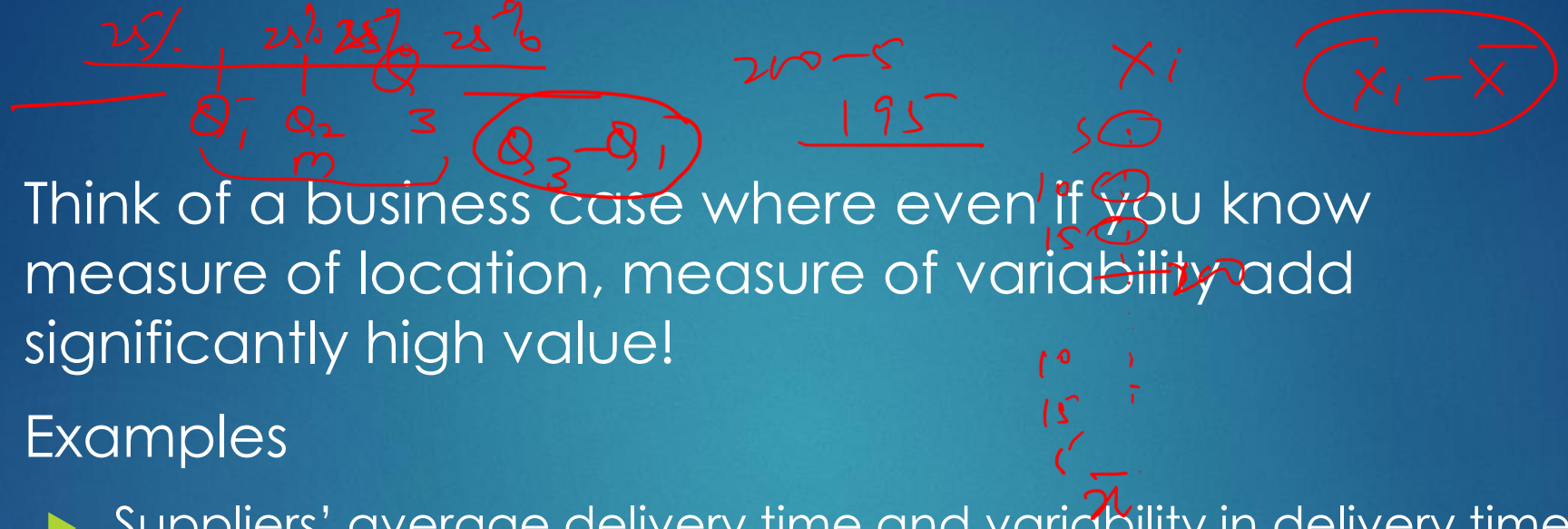
Measures of Central Tendency: Quartiles



- First Quartile (Q_1) = 25th Percentile
- Second Quartile (Q_2) = 50th Percentile
- Third Quartile (Q_3) = 75th 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, Variance, 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
  - Measure of variability that utilizes all the data
  - average of the squared deviations between each data value and the mean

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

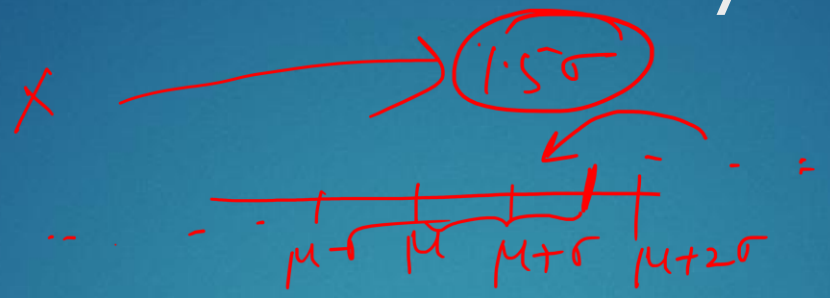
Measures of Variability (Dispersion)

- 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

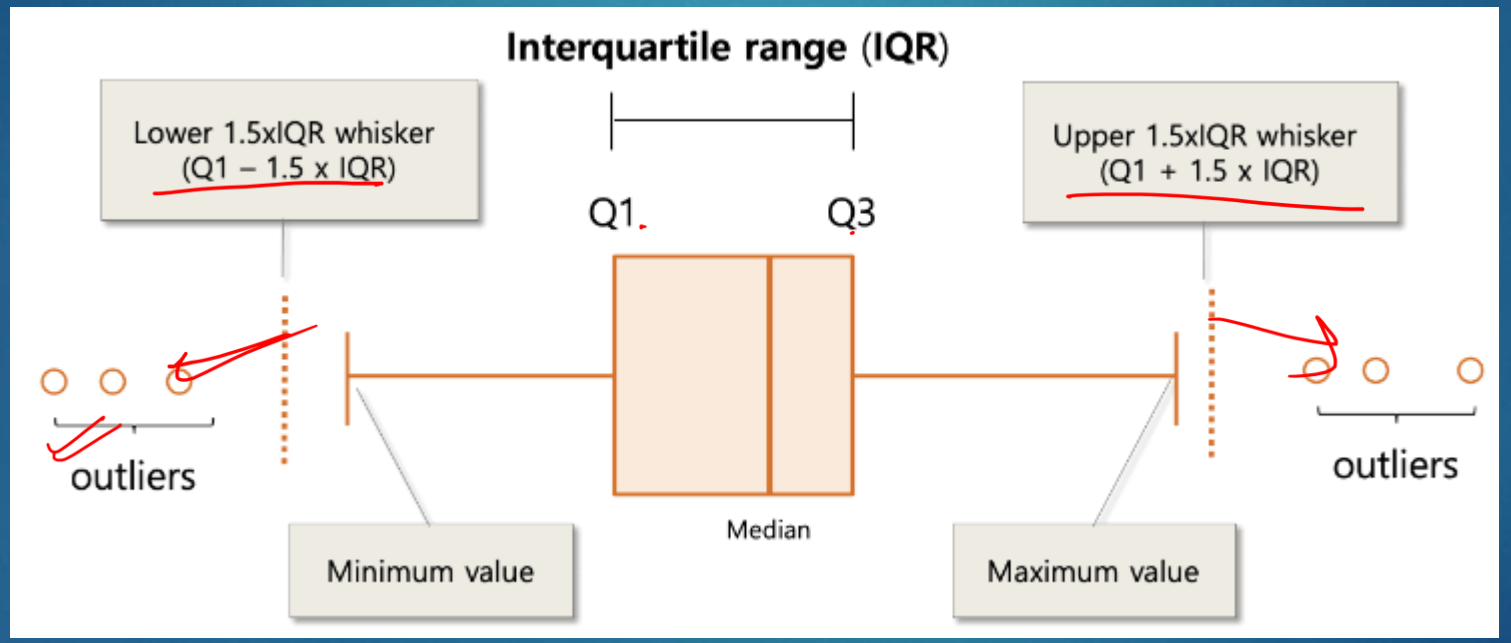
Five-Number Summaries & Box Plots

- 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


Interquartile range (IQR)
Lower 1.5xIQR whisker
(Q1 – 1.5 x IQR)
Upper 1.5xIQR whisker
(Q1 + 1.5 x IQR)
Q1
Q3
outliers
outliers
Median
Minimum value
Maximum value

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: $r_{xy} = \frac{s_{xy}}{s_x s_y}$