BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods(ISM)

Slides Modified by
Dr YVK Ravi Kumar
yvk.ravikumar@pilani.bits-pilani.ac.in

- Section:
- Faculty Name & Mail ID: Prof. Mandar Vijay Datar

(mandar.datar@wilp.bits-Pilani.ac.in)

Session 1

(27th July 2025)

Overview of the course & Basic of Statistics

Overview of the course

- **Module 1** : Basic Probability & Statistics
- **Module 2** : Conditional Probability & Bayes’ theorem
- **Module 3** : Probability Distributions
- **Module 4** : Hypothesis Testing
- **Module 5** : Prediction & Forecasting
- **Module 6** : Prediction & Forecasting; Gaussian Mixture model &Expectation Maximization

TEXT BOOKS

- T1 : Statistics for Data Scientists, An introduction to Probability, Statistics and Data Analysis, Maurits Kaptein et al, Springer 2022
- T2 : Probability and Statistics for Engineering and Sciences,8th Edition, Jay L Devore, Cengage Learning
- T3 : Introduction to Time Series and Forecasting, Second Edition , Peter J Brockwell, Richard A Davis, Springer.

Evaluation Components

	Name	Type	Weightage
EC 1	Quiz 1 & Quiz 2	Online	10%
	Assignment 1 & Assignment 2	Online	20%
EC 2	Mid semester Exam	Closed Book	30%
EC 3	Comprehensive Exam	Open Book	40%

Module 1: (Basic Probability & Statistics)

Contact Session	List of Topic Title	Reference
CS - 1	Measures of Central Tendency & Measures of Variability, Data – Symmetric & Asymmetric, outlier detection, 5 point summary	T1 & T2

Statistics

➢ **Statistics may be defined as science that is employed to**

- Collect the data
- Present and organize the data in a systematic manner
- Analyse the data
- Infer about the data
- Take decision from the data.

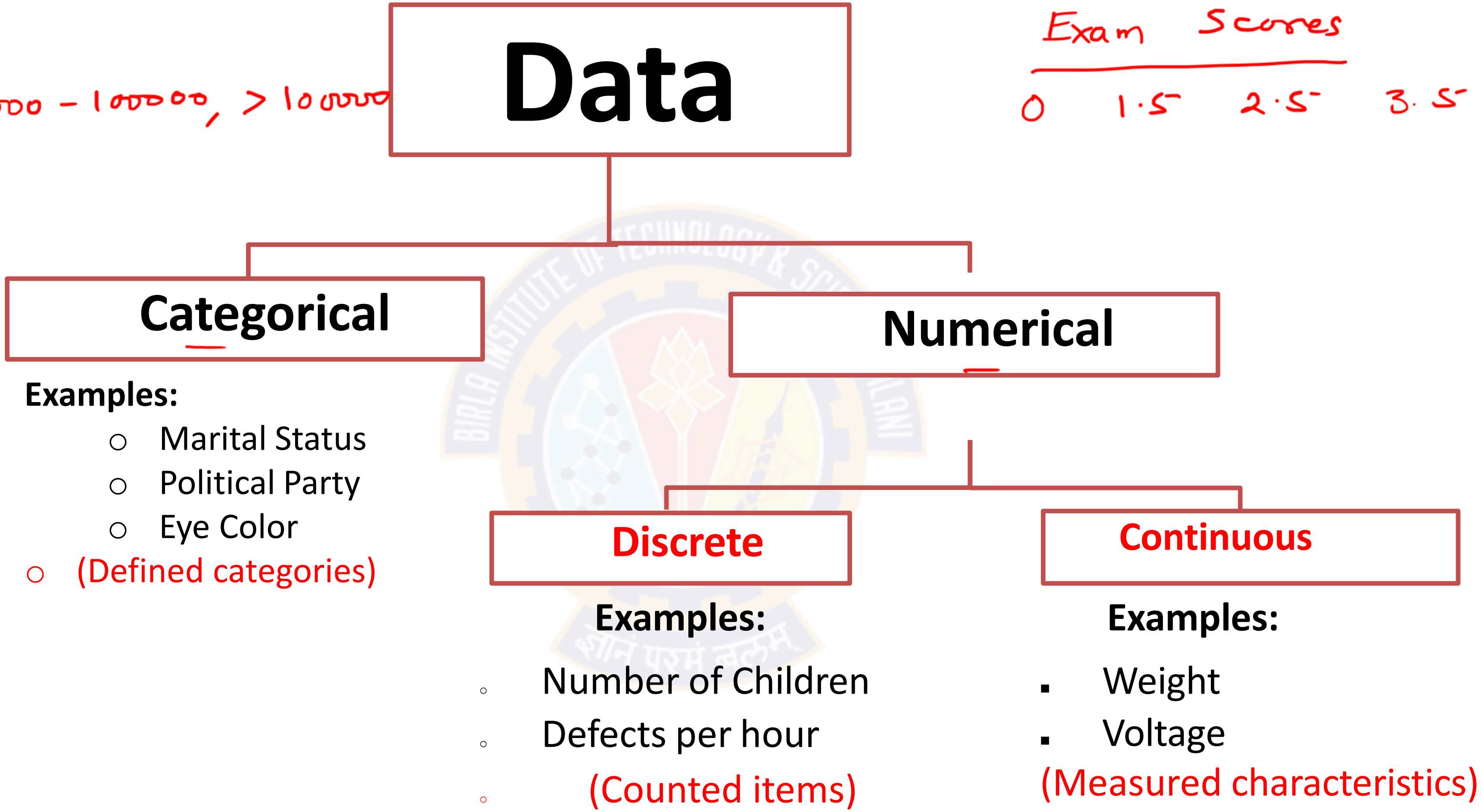
Data - Types

Salary

< 50000, 50000 - 100000, > 100000

Exam Scores

0 1.5 2.5 3.5 4.5



Data

Categorical

Examples:

- Marital Status
- Political Party
- Eye Color
- (Defined categories)

Numerical

Discrete

Examples:

- Number of Children
- Defects per hour
- (Counted items)

Continuous

Examples:

- Weight
- Voltage

(Measured characteristics)

Levels of Data Measurement

- Nominal — Lowest level of measurement
- Ordinal
- Interval
- Ratio — Highest level of measurement

✓ categorical

❑ **Nominal**
- ○ A **nominal scale** classifies data into distinct categories in which no ranking is implied
- ○ Example : Gender, Marital Status, voting to different party

✓ ❑ **Ordinal scale**
- ○ An **ordinal scale** classifies data into distinct categories in which ranking is implied
- ○ Example:

1. Product satisfaction: Satisfied, Neutral, Unsatisfied
2. Faculty rank: Professor, Associate Professor, Assistant Professor
3. Student Grades: A, B, C, D, F
4. Medals won: Gold, Silver, Bronze

$40^\circ C = 2 \times 20^\circ C$

Numerical

✓ ❑ **Interval scale**
- ○ An **interval scale** is an ordered scale in which the difference between measurements is a meaningful quantity but the measurements do not have a true zero point.
- ○ **Example**

1. Temperature in Fahrenheit and Celsius
2. Months of the Year: there's no month called zero and we can't say January is twice as much as June.

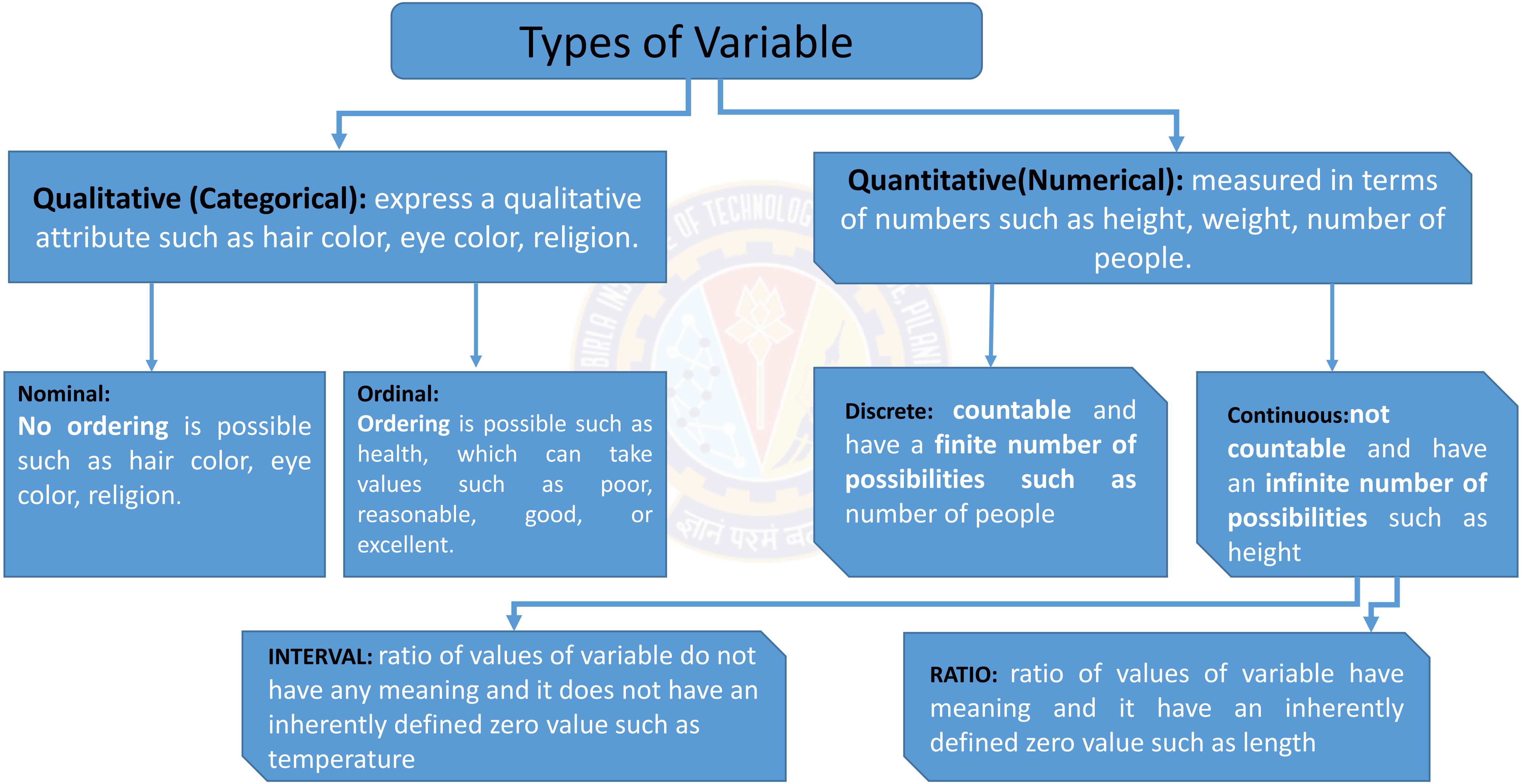
Jan = 1 June = 6

$1 \times 6 = 6$

✓ ❑ **Ratio Scale**
- ○ A **ratio scale** is an ordered scale in which the difference between the measurements is a meaningful quantity and the measurements have a true zero point.
- ○ Examples

Weight , Age , Salary

A's weight 20 kg. B's weight = 40 kg



Types of Variable

Qualitative (Categorical): express a qualitative attribute such as hair color, eye color, religion.

Nominal:
No ordering is possible such as hair color, eye color, religion.

Ordinal:
Ordering is possible such as health, which can take values such as poor, reasonable, good, or excellent.

Quantitative(Numerical): measured in terms of numbers such as height, weight, number of people.

Discrete: **countable** and have a **finite number of possibilities such as** number of people

Continuous:not countable and have an **infinite number of possibilities** such as height

INTERVAL: ratio of values of variable do not have any meaning and it does not have an inherently defined zero value such as temperature

RATIO: ratio of values of variable have meaning and it have an inherently defined zero value such as length

Example

Question	Type of Variable
Types of vehicle owned Two wheeler, four wheeler	Categorical : Nominal
Product satisfaction Unsatisfied, neutral, fairly satisfied, satisfied	Categorical : Ordinal
To how many magazines do you currently subscribed Zero, One, Two, Three, Four	Discrete
How tall are you (in inches)	Continuous: Ratio
Weight (in Kilograms)	Continuous: Ratio
Temperature (in degrees Celsius or degrees Fahrenheit)	Continuous: Interval

Measures of Central Tendency

- Measure of central tendency provides a very convenient way of describing a set of scores with a single number that describes the **PERFORMANCE** of the group.

- Also defined as a single value that is used to describe the “**center**” of the data.

- Three commonly used measures of central tendency:
  1. Mean
  2. Median
  3. Mode

Mean

- ➢Also referred as the "**arithmetic average**"
- ➢The most commonly used measure of the center of data
- ➢Numbers that describe what is average or typical of the distribution
- ➢Computation of Sample Mean:

✓ $$\bar{Y} = \frac{\sum Y}{N}$$ = Σ Y/ N = (Y1 + Y2 + Y3 + … Yn) / N where

- "Y bar" equals the sum of all the scores, Y, divided by the number of scores, N.

- ➢Computation of the Mean for grouped Data

✓ ➢ $$\bar{Y} = \frac{\sum f\,Y}{N}$$ where f Y = a score multiplied by its frequency

$= \frac{57}{11} \leftarrow \frac{\Sigma fy}{\Sigma f}$

$x = 10, 12, 15, 20, 21, 24$

$\bar{x} = \frac{10+12+15+20+21+24}{6}$

$x = 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 8$

x	f	$x \cdot f$
2	1	2
3	1	3
4	2	8
5	3	15
6	1	6
7	1	7
8	2	16
	$\Sigma f = 11 = N$	57

Mean: Ungrouped Data

- Suppose you define the time to get ready as the time (rounded to the nearest minute) from when you get out of bed to when you leave your home. You collect the times shown below for 10 consecutive work days:

Day	1	2	3	4	5	6	7	8	9	10
Time (min)	39	29	43	52	39	44	40	31	44	35

$$\overline{X} = \frac{\text{Sum of the values}}{\text{Number of values}}$$

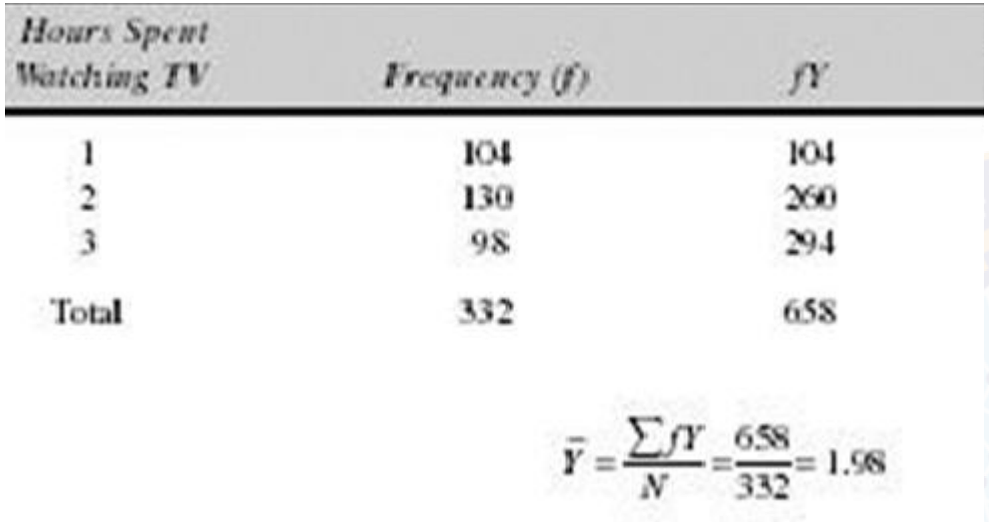
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\overline{X} = \frac{39 + 29 + 43 + 52 + 39 + 44 + 40 + 31 + 44 + 35}{10}$$

$$\overline{X} = \frac{396}{10} = 39.6$$

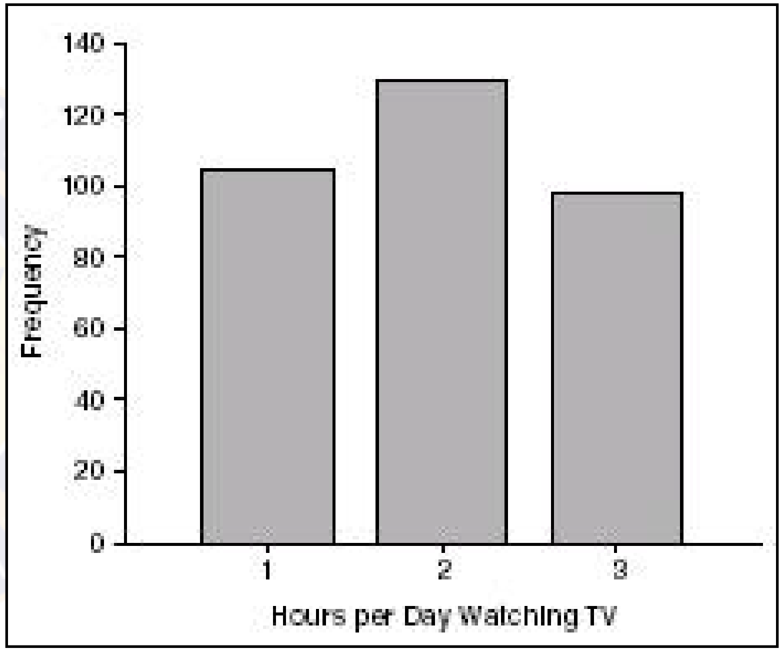
Mean: Grouped Scores

Data of Children watching TV in Bengaluru



Hours Spent Watching TV	*Frequency (f)*	*fY*
1	104	104
2	130	260
3	98	294
Total	332	658

$$\bar{Y} = \frac{\sum fY}{N} = \frac{658}{332} = 1.98$$



Mean

- **PROPERTIES**
  - It measures stability. Mean is the most stable among other measures of central tendency because every score contributes to the value of the mean.
  - It may easily affected by the extreme scores. outliers
  - The sum of each score's distance from the mean is zero.
  - It can be applied to interval level of measurement
  - It may not be an actual score in the distribution
  - It is very easy to compute.
- **When to Use the Mean**
  - Other measures are to be computed such as standard deviation, coefficient of variation and skewness
  - Sampling stability is desired.

Batsman A

Avg = 30

Company A = 29, 30, 31

Batsman = B

Company B = 20, 30, 40

Avg = 30

Company C

10, 20, 500

The Mode

- The category or score with the largest frequency (or percentage) in the distribution.
- The mode can be calculated for variables with levels of measurement that are: nominal, ordinal, or interval-ratio.

Example:

- A systems manager in charge of a company s network keeps track of the number of server failures that occur in a day. Compute the mode for the following data, which represents the number of server failures in a day for the past two weeks:.

1 3 ✓ 0 3 ✓ 26 2 7 4 2 3 ✓ 3 ✓ 6 3 ✓

- Because 3 appears five times, more times than any other value, the mode is 3. Thus, the systems manager can say that the most common occurrence is having three server failures in a day.

Multimodal = Having more than one mode

Mode

Properties

- It can be used when the data are qualitative ✓ as well as quantitative.
- It may not be unique. ✓
- It is not affected by extreme values. ✓

When to Use the Mode

- When the data set is measured on a nominal scale

<u>Satisfactory</u> <u>Question</u>

The Median

- The score that divides the distribution into two equal parts, so that half the cases are above it and half below it.
- You compute the median value by following one of two rules:

○ ***Rule 1*** If there are an *odd* number of values in the data set, the median is the middle-ranked value.

○ ***Rule 2*** If there are an *even* number of values in the data set, then the median is the *average* of the two middle ranked values.

- The median is the middle score, or average of middle scores in a distribution.
  - Fifty percent (50%) lies below the median value and 50% lies above the median value.
  - It is also known as the middle score or the 50th percentile.

$2, 5, \boxed{6}, 7, 8$

$n = 5$

$\frac{n+1}{2}$ ✓

$4, 6, \boxed{7}\boxed{9}, 11, 12$

$8 = \text{median}$

$\left(\frac{n}{2}\right)^{th}, \left(\frac{n}{2}\right)^{th}+1 \rightarrow \text{Average}$

The Median

- **Example-1**: The three-year annualized returns for the seven small-cap growth funds with low risk are ranked from the smallest to the largest:

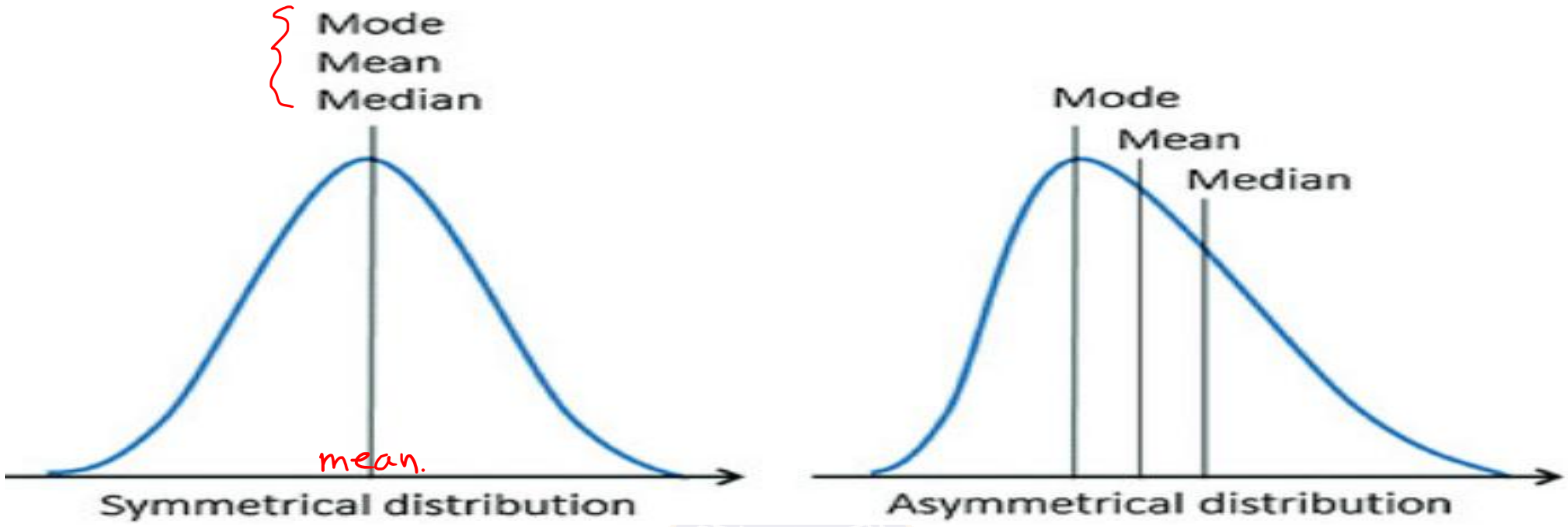
  The Median is the 4th position value i.e. 22.4.

- **Example-2**: Consider the time taken for reaching office from home in minutes arranged from lowest to highest

29 31 35 39 39 40 43 44 44 52

- The number of observations are 10. Hence the median is the average of values at the 5th and 6th Position, i.e. average of 39 and 40 = **39.5**

Data : Symmetrical and Asymmetrical



- Central tendency Median is used for both Symmetrical and Asymmetrical data
- While Mean or Mode or Median can be used as central tendency for Symmetrical data.

Data distribution

- Skewed

○ Negatively (Left):mean < median

  - Distributions with a left skew have long left tails;

○ Positively (Right) : mean > median

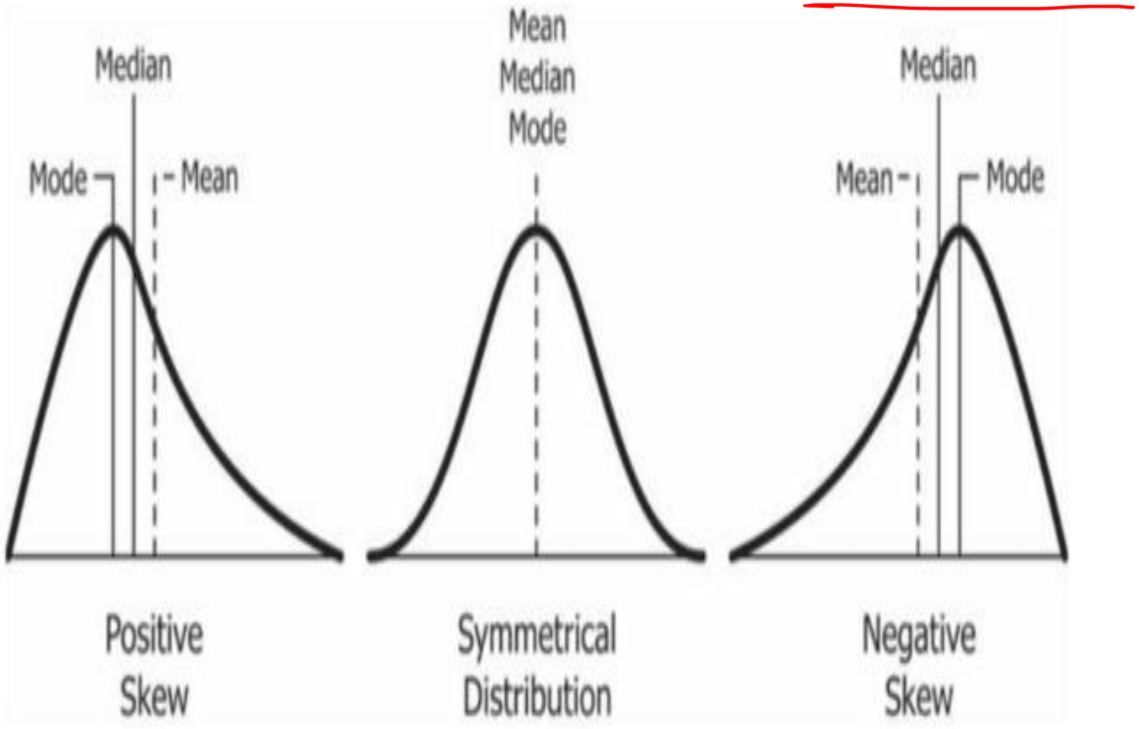
  - Distributions with a right skew have long right tails.

- Bimodal : has two distinct modes

- Multi-modal : has more than 2 distinct modes

Symmetrical.

Left skewed



Note: In science, an empirical relationship is a relationship that is supported by experiment and observation but not necessarily supported by theory. For moderately skewed data (Skewness between -0.5 to 0.5) empirical relationship is 3Medain= Mode+2Mean

Why mean is not enough?

$2, 2, 3, 5, \boxed{5}, 5, 7, 8, 8$

$\frac{9+1}{2} = 5^{th}$ pos

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

Statistical measures	Group 1
Mean	5 ✓
Median	5 ✓
Mode	5 ✓

Sl. No.	X_2
1	1
2	15
3	5
4	5
5	6
6	3
7	5
8	2
9	3
Total	45

Statistical measures	Group 2
Mean	5
Median	5
Mode	5

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

Statistical measures	Group 1 & 2
Mean	5
Median	5
Mode	5

Do we need any other measure?

Answer: Yes

Avg = 50 49, 50, 51 → 'A'

Avg = 50 0, 50, 100 → 'B'

Measures of variability

- Three Measures of Variability:
  - The Range → Batsman A [49, 51]
    Batsman B [0, 100]
  - The Variance
  - The Standard Deviations

S.D. = $\sqrt{Var.}$

Measure of Variability

➢Variability can be defined several ways:

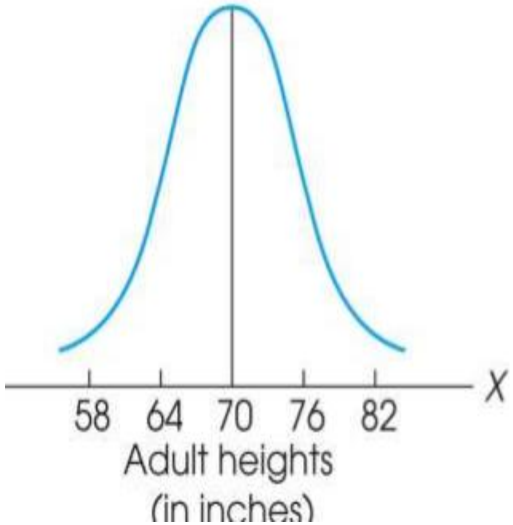
o A quantitative distance measure based on the differences between scores

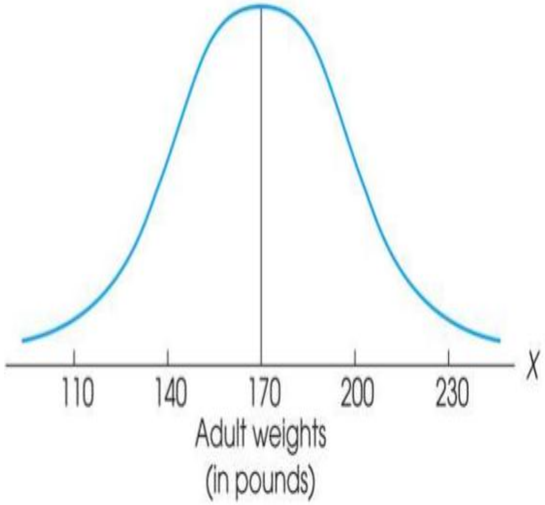
o Describes distance of the spread of scores or distance of a score from the mean

➢Purposes of Measure of Variability:

o Describe the distribution

o Measure how well an individual score represents the distribution





The Ranges

- The distance covered by the scores in a distribution – From smallest value to highest value
- For continuous data, real limits are used

$$\text{Range} = \text{URL for } X_{max} - \text{LRL for } X_{min}$$

- Based on two scores, not all the data – An imprecise, unreliable measure of variability

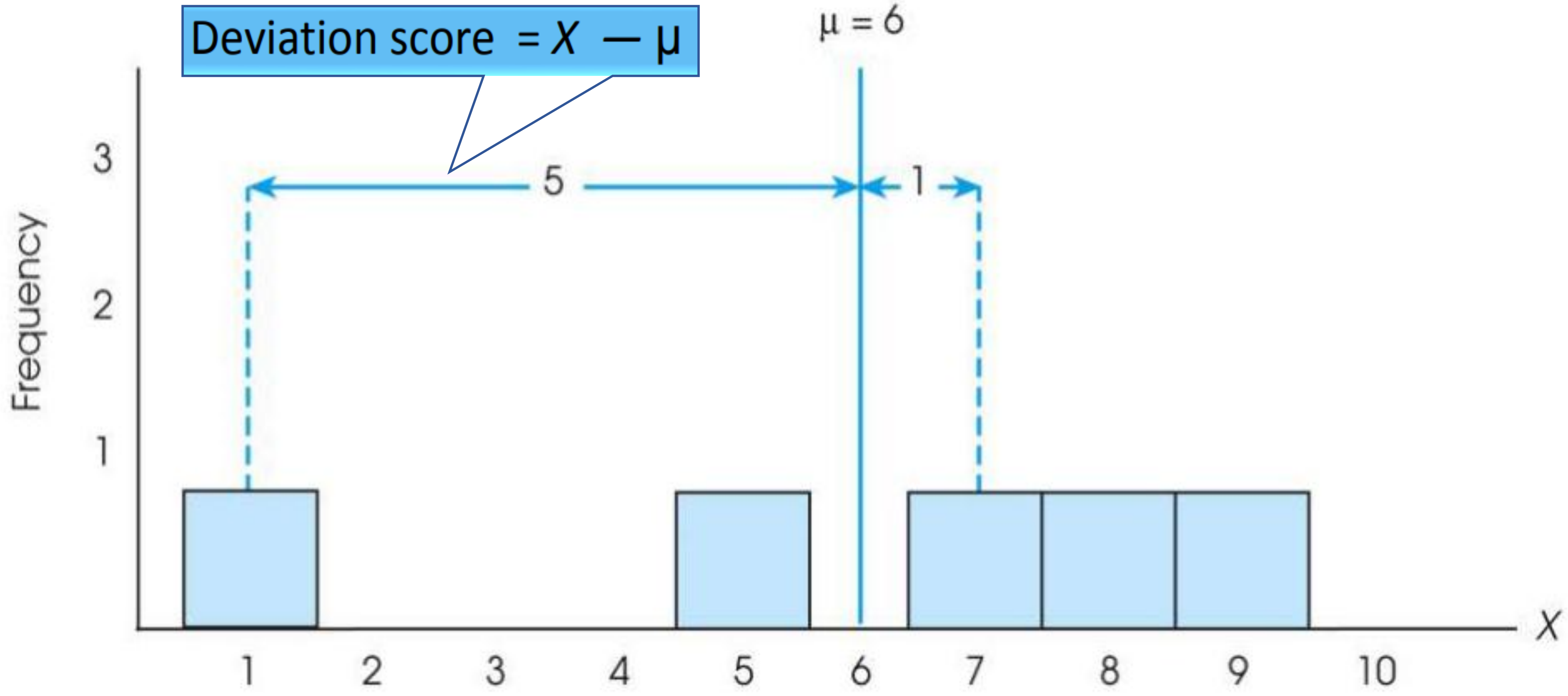
Example: For a set of scores: 7, 2, 7, 6, 5, 6, 2

Range = Highest Score minus Lowest score = 7 - 2 = 5

The Standard Deviation

- Most common and most important measure of variability is the standard deviation
  - A measure of the standard, or average, distance from the mean
  - Describes whether the scores are clustered closely around the mean or are widely scattered
- Calculation differs for population and samples
- Variance is a necessary *companion concept* to standard deviation but *not the same* concept

The Standard Deviation


Deviation score $= X - \mu$

Exercise : Find out the deviations of all the data points with the mean....and then find the 'mean deviation'.

The Standard Deviation

- Mean deviations will always be ‘zero’ ! (because Mean is a balance point)

- Then, how do you find ‘Standard Deviation’ ?

Need a new strategy

The Standard Deviation

- New Strategy :
  a) First square each deviation score
  b) Then sum the Squared Deviations (SS)
  c) Average the squared deviations

- Mean Squared Deviation is known as **“Variance”**
- Variability is now measured in squared units

$$\boxed{Standard\ Deviation = \sqrt{Variance}}$$

The Variance

Variance equals mean (average) squared deviation (distance) of the scores from the mean

$$\text{Variance} = \frac{\text{sum of } \underline{\text{squared deviations}}}{\text{number of scores}}$$

where $SS = \sum(\underline{X - \mu})^2$

The Population Variance

- Population variance equals mean (average) squared deviation (distance) of the scores from the population mean

- Variance is the average of squared deviations, so we identify population variance with a lowercase Greek letter sigma squared: σ^2

- Standard deviation is the square root of the variance, so we identify it with a lowercase Greek letter sigma: σ

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

Statistical measures	Group 1 & 2
Mean	5
Median	5
Mode	5

$$\overline{X} = \frac{\sum_{1}^{9} x_i}{N} = \frac{45}{9} = 5$$ ✓

$$S = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{n-1}}$$

$$S = \sqrt{\frac{44}{8}} = \underline{\underline{2.345}}$$

Variance → Population $\frac{\Sigma(x-\bar{x})^2}{N}$

Variance → Sample $\frac{\Sigma(x-\bar{x})^2}{n-1}$

For x_1 $\quad \underline{\mu = 5}$ $\quad \bar{x} = 5$

	X_1	$X-5$	$(X-5)^2$
1	2	-3	9
2	8	3	9
3	5	0	0
4	3	-2	4
5	7	2	4
6	8	3	9
7	5	0	0
8	2	-3	9
9	5	0	0
		0	44

$$\text{Var} = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{44}{8} = 5.5$$

$$S.D = \sqrt{5.5} =$$

Second Dataset

Sl. No.	X_2
1	1
2	15
3	5
4	5
5	6
6	3
7	5
8	2
9	3
Total	45

$$\overline{X} = \frac{\sum_{1}^{9} x_i}{N} = \frac{45}{9} = 5$$

$$S = \sqrt{\frac{\sum \left(X - \overline{X}\right)^2}{n - 1}}$$

$$S = \sqrt{\frac{134}{8}} = 4.093$$ ✓

Standard Deviation and Variance for a Sample

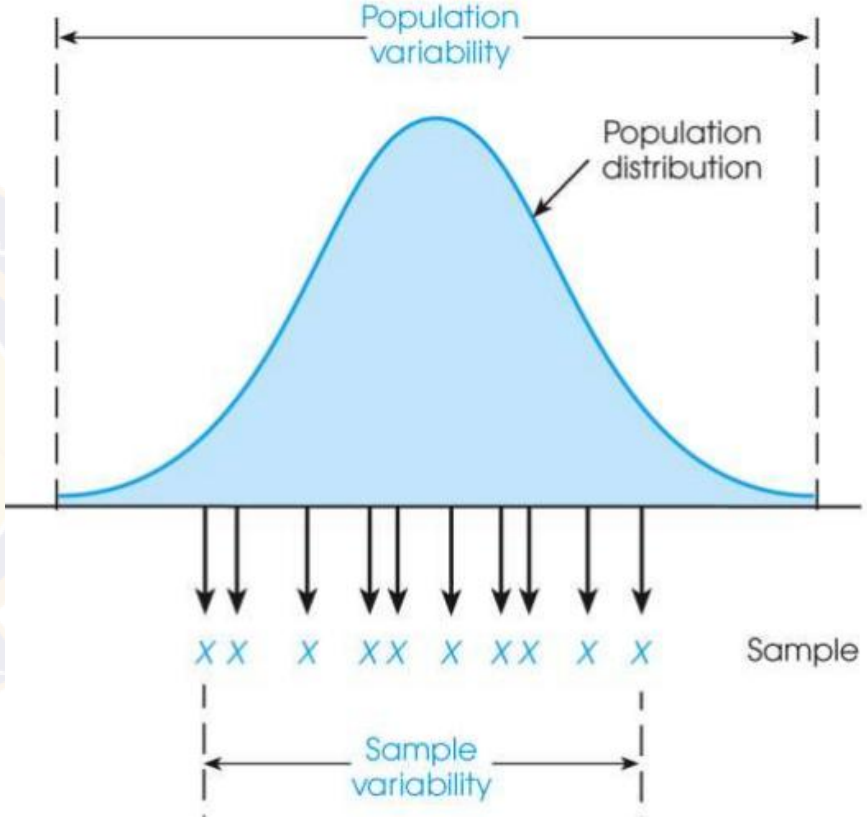
- Goal of inferential statistics:
  - Draw general conclusions about population based on limited information from a sample
- Samples differ from the population
  - Samples have less variability
  - Computing the Variance and Standard Deviation in the same way as for a population would give a biased estimate of the population values

Sample Standard Deviation and Variance

- Sum of Squares (SS) is computed as before
- Formula for Variance has n-1 rather than N in the denominator
- Notation uses s instead of σ

$$\text{variance of sample} = s^2 = \frac{SS}{n-1}$$

$$\text{standard deviation of sample} = s = \sqrt{\frac{SS}{n-1}}$$



Degrees of Freedom

- Population variance
  - Mean is known
  - Deviations are computed from a known mean

- Sample variance as estimate of population
  - Population mean is unknown
  - Using sample mean restricts variability

- Degrees of freedom
  - Number of scores in sample that are independent and free to vary
  - Degrees of freedom (df) = n – 1

Learning Check

Select the correct option

a) A sample of four scores has SS = 24. What is the variance?

(1) The variance is 6

(2) The variance is 7

(3) The variance is 8

(4) The variance is 12

b) A sample systematically has less variability than a population

True / False ?

b) The standard deviation is the distance from the Mean to the farthest point on the distribution curve

True / False ?

Learning Check

Select the correct option

a) A sample of four scores has SS = 24. What is the variance?

(1) The variance is 6

(2) The variance is 7

(3) The variance is 8

(4) The variance is 12

b) A sample systematically has less variability than a population

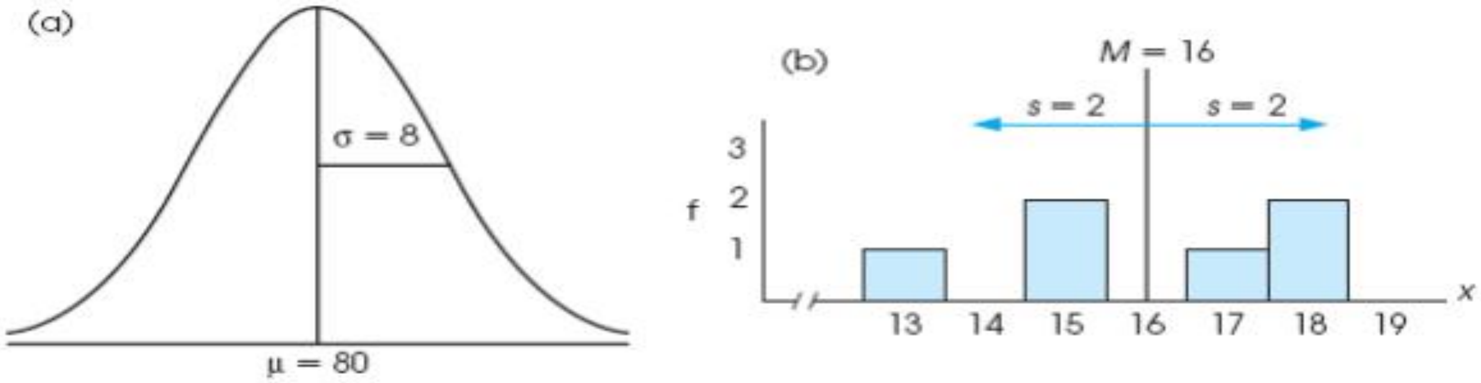
True

b) The standard deviation is the distance from the Mean to the farthest point on the distribution curve

False

Descriptive Statistics

- A standard deviation describes scores in terms of distance from the mean
- Describe an entire distribution with just two numbers (M and s)
- Reference to both allows reconstruction of the measurement scale from just these two numbers
- Means and standard deviations together provide extremely useful descriptive statistics for characterizing distributions

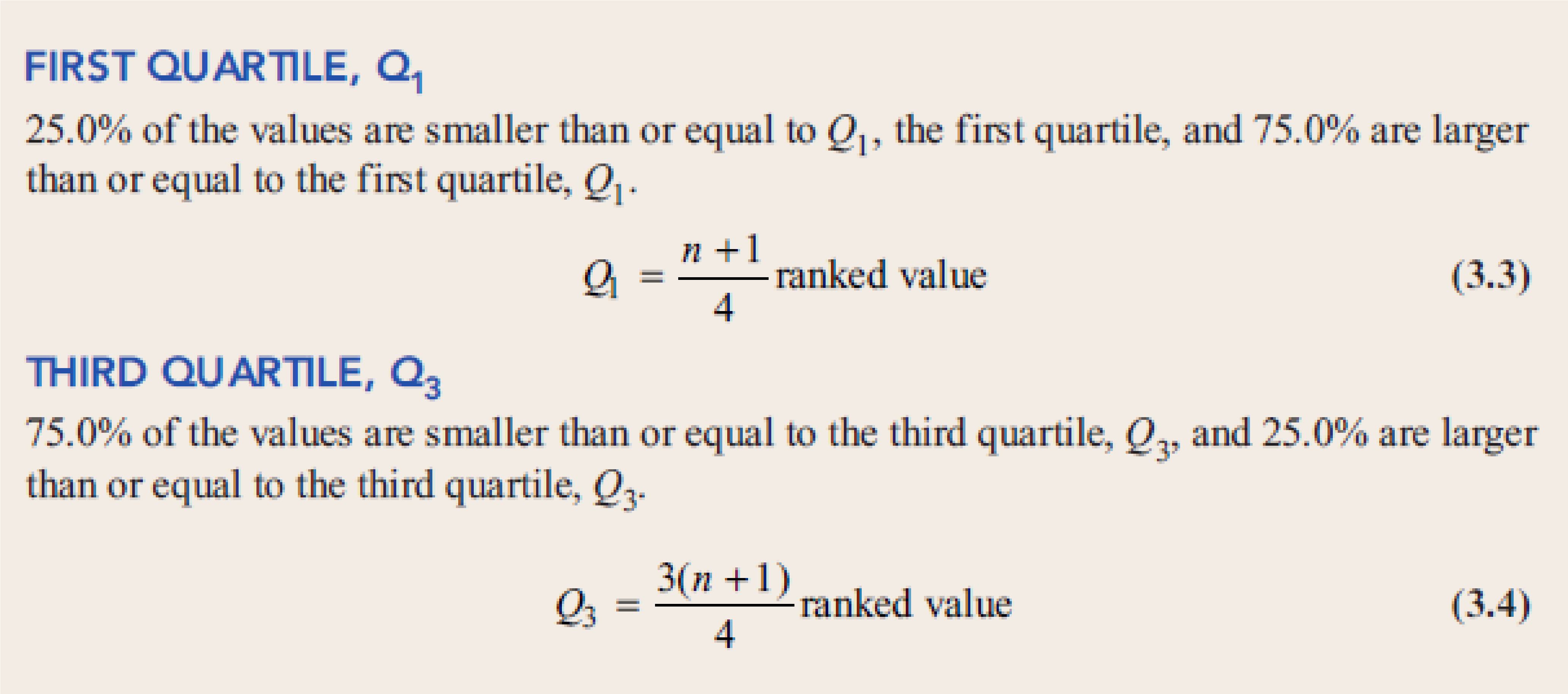


Five point summary of Data

The five number summary of data includes 5 items:

- **Minimum**.
- **Q1** (the first quartile, or the 25% mark).
- **Median**.
- **Q3** (the third quartile, or the 75% mark).
- **Maximum**.

Quirtile



FIRST QUARTILE, Q_1

25.0% of the values are smaller than or equal to Q_1, the first quartile, and 75.0% are larger than or equal to the first quartile, Q_1.

$$Q_1 = \frac{n+1}{4} \text{ ranked value} \qquad (3.3)$$

THIRD QUARTILE, Q_3

75.0% of the values are smaller than or equal to the third quartile, Q_3, and 25.0% are larger than or equal to the third quartile, Q_3.

$$Q_3 = \frac{3(n+1)}{4} \text{ ranked value} \qquad (3.4)$$

Rules to identify Quartiles

Rule 1 If the result is a whole number, then the quartile is equal to that ranked value. For example, if the sample size $n = 7$, the first quartile, $Q1$, is equal to the $(7 + 1)/4 =$ second ranked value.

Rule 2 If the result is a fractional half (2.5, 4.5, etc.), then the quartile is equal to the average of the corresponding ranked values. For example, if the sample size $n = 9$, the first quartile, $Q1$, is equal to the

$(9 + 1)/4 = 2.5$ ranked value, halfway between the second ranked value and the third ranked value.

Rule 3 If the result is neither a whole number nor a fractional half, you round the result to the nearest integer and select that ranked value. For example, if the sample size $n = 10$, the first quartile, $Q1$, is equal to the $(10 + 1)/4 = 2.75$ ranked value. Round 2.75 to 3 and use the third ranked value.

Interquartile range (IQR)

- It is measure of Variation
- Also Known as Mid-spread : Spread in the Middle 50%
- Difference Between Third & First Quartiles:
- Not Affected by Extreme Values
- Interquartile Range = IQR = $Q_3 - Q_1$

$n = 9$ $\quad Q_1 = \frac{n+1}{4} = 2.5^{th}$ pos

12.5

$Q_2 = \frac{9+1}{2} = 5^{th}$ pos

$Q_2 = 16$

Data in Ordered Array: 11 12 ↑ 13 16 16 17 17 ↑ 18 21

$Q_3 = \frac{3(n+1)}{4} = 7.5^{th}$ pos.

$$\textbf{Position of } Q_1 = \frac{1\bullet(9+1)}{4} = 2.50, \qquad Q_1 = 12.5$$

$$\textbf{Position of } Q_3 = \frac{3\bullet(9+1)}{4} = 7.50, \qquad Q_3 = 17.5$$

$Q_3 = 17.5$

min = 11

$Q_1 = \underline{\underline{12.5}}$

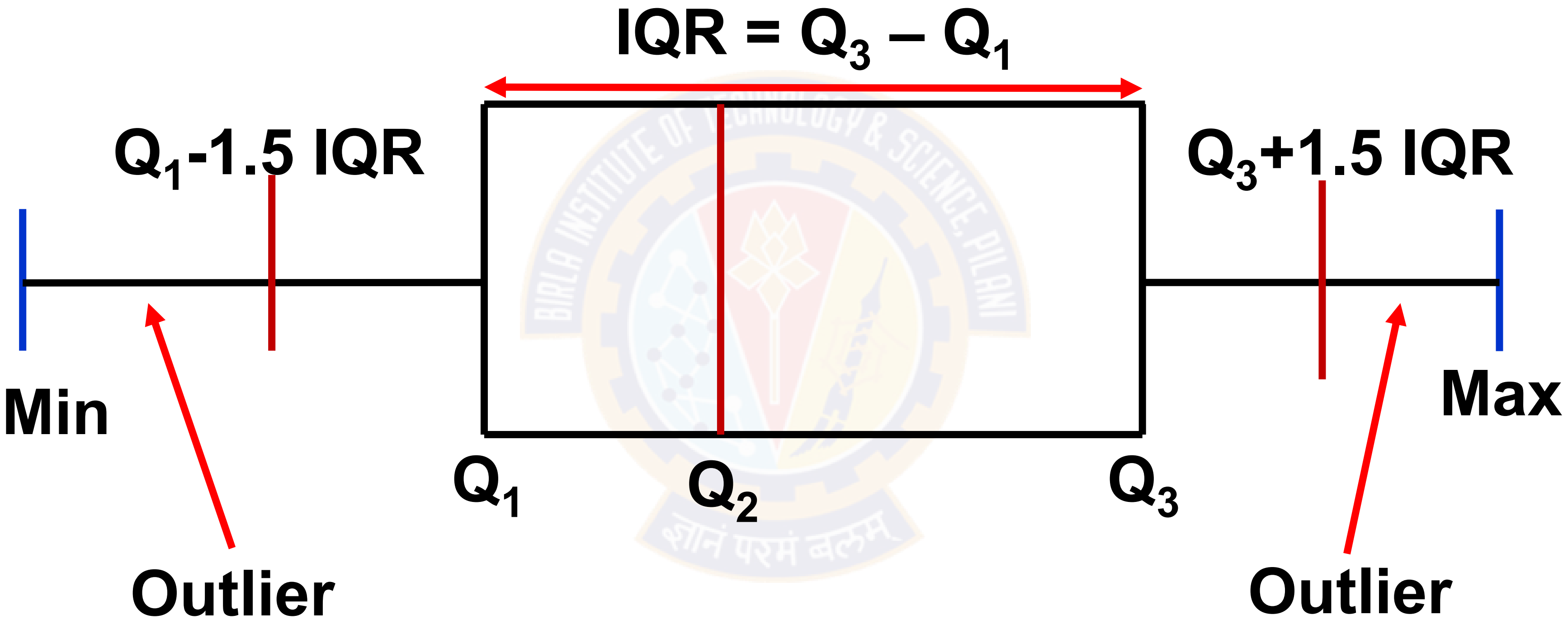
$Q_2 = 16$

$Q_3 = 17.5$

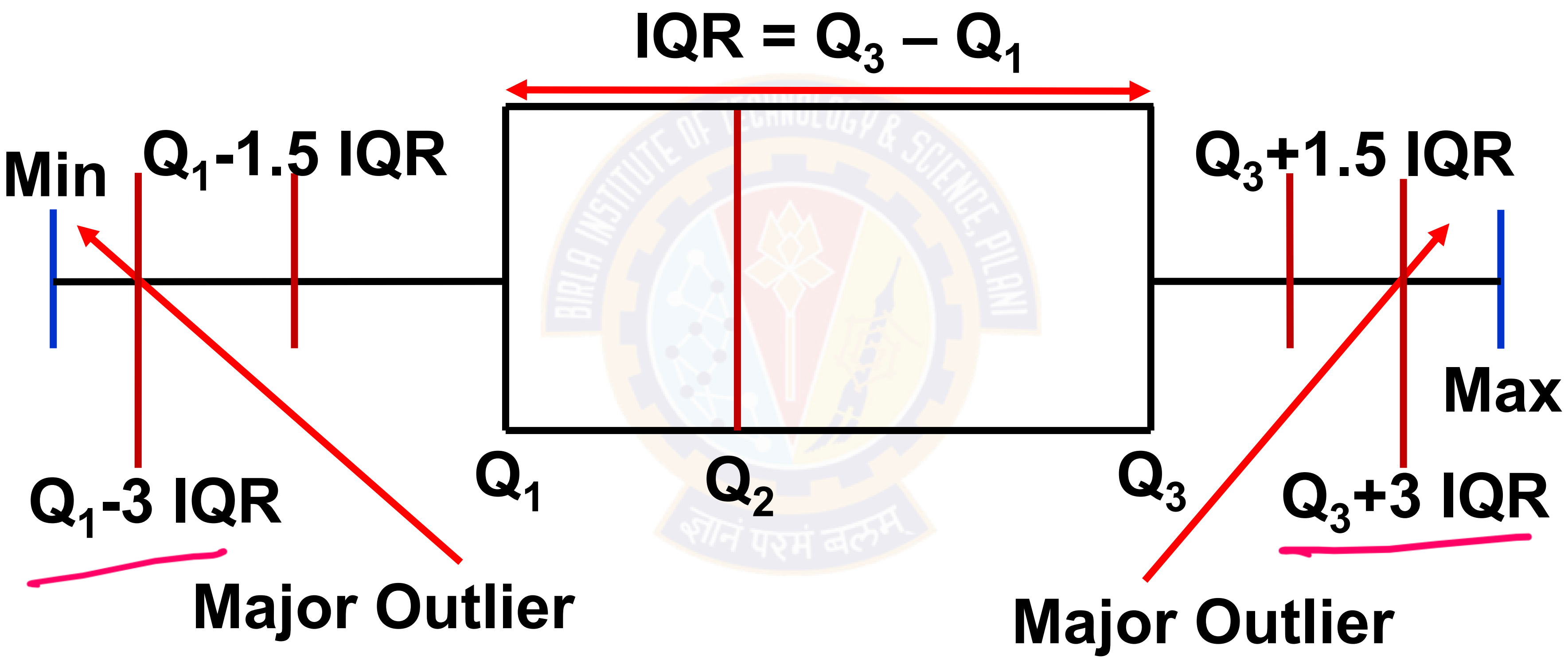
max = 21

Interquartile Range = IQR= $Q_3 - Q_1$ = 17.5 - 12.5 = 5

Box and Whisker plot



Box and Whisker plot



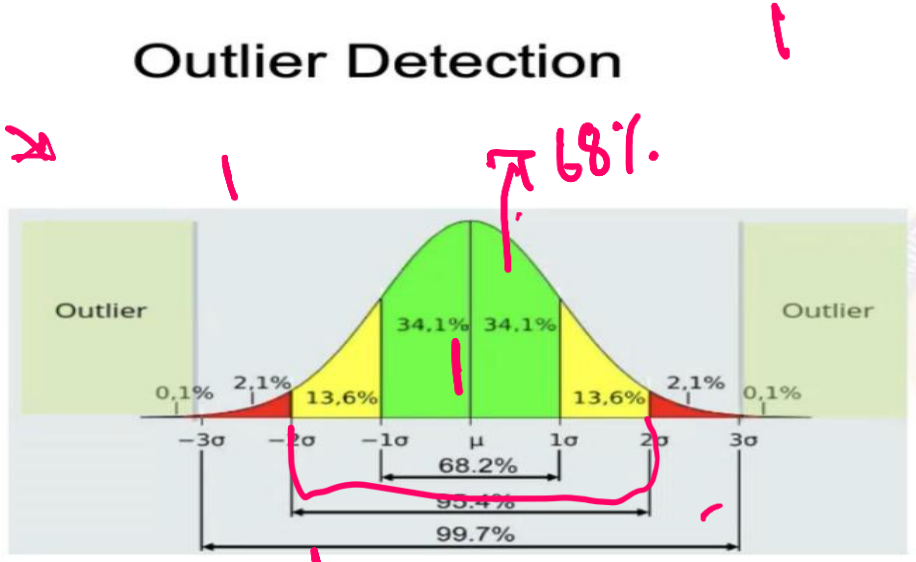
Potential outliers

- The lower limit and upper limit of a data set are given by:

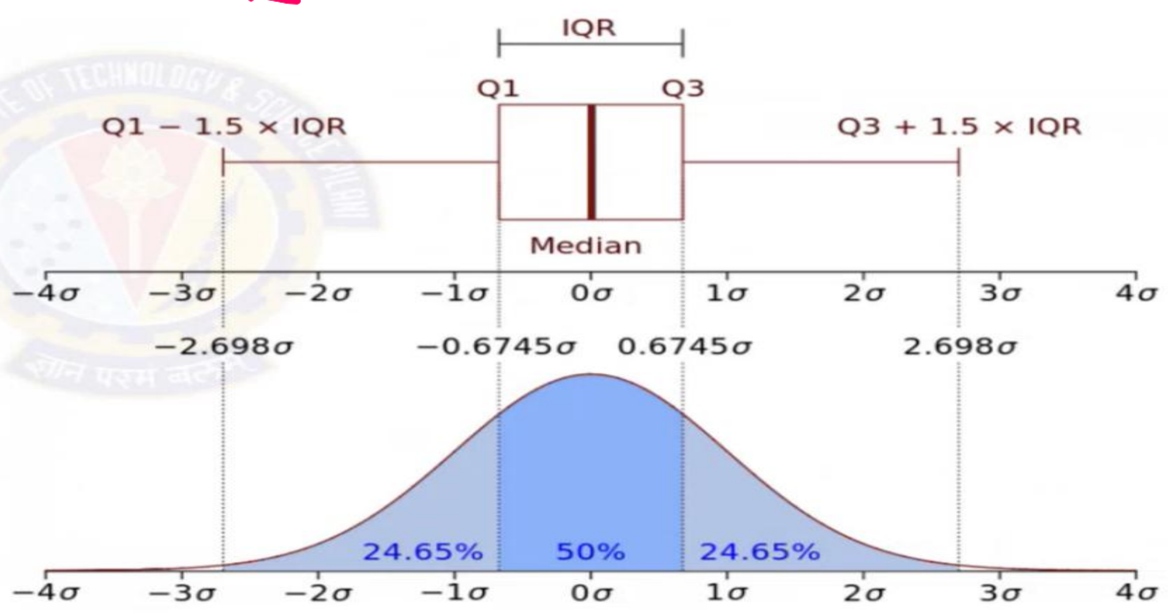
  Lower limit = Q_1 - 1.5 x IQR

  Upper limit = Q_3 + 1.5 x IQR

- Data points that lie below the lower limit or above the upper limit are potential outliers.



mean ± 3σ



Summary

1.**Measures of Central Tendency:** Mean, Median, Mode

2. **Measures of Variability**: Range, Standard Deviation , Variance

3. Symmetric and Asymmetric distribution

4. Five point summary

5. Outliers

Practice Problem:

- Q.1 A sample of 77 individuals working at a particular office was selected and the noise level (dBA) experienced by everyone is the following data:

55.3, 55.3, 55.3, 55.9, 55.9, 55.9, 55.9, 56.1, 56.1, 56.1, 56.1, 56.1, 56.1,
56.8, 56.8, 57.0, 57.0, 57.0, 57.8, 57.8, 57.8, 57.9, 57.9, 57.9, 58.8, 58.8,
58.8, 59.8, 59.8, 59.8, 62.2, 62.2, 63.8, 63.8, 63.8, 63.9, 63.9, 63.9, 64.7,
64.7, 64.7, 65.1, 65.1, 65.1, 65.3, 65.3, 65.3, 65.3, 67.4, 67.4, 67.4, 67.4,
68.7, 68.7, 68.7, 68.7, 69.0, 70.4, 70.4, 71.2, 71.2, 71.2, 73.0, 73.0, 73.1,
73.1, 74.6, 74.6, 74.6, 74.6, 79.3, 79.3, 79.3, 79.3, 83.0, 83.0, 83.0. 92.4

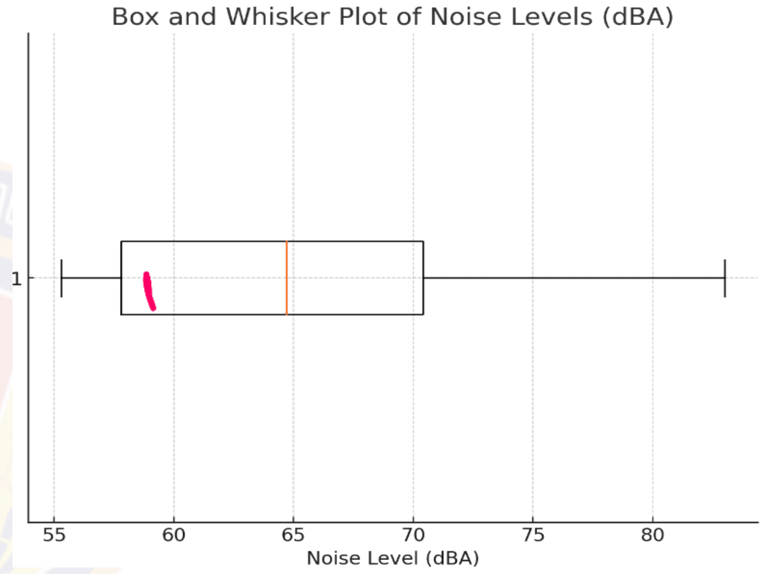
Find a) Arithmetic Mean, SD, variance, and IQR

b) Draw box and whisker plot

c) Comment on the outliers, if any.

Discussion

- **Arithmetic Mean:** 64.89 dBA
- **Standard Deviation (SD):** 7.80 dBA
- **Variance:** 60.88 $(\text{dBA})^2$
- **Interquartile Range (IQR):** 12.60 dBA



- **Lower Bound:** $Q1 - 1.5 \times IQR = \overset{57.8}{\boxed{59.8}} - (1.5 \times 12.6) = \overline{\underline{40.9}}$
- **Upper Bound:** $Q3 + 1.5 \times IQR = \underset{70.4}{\underline{72.4}} + (1.5 \times 12.6) = \underline{91.3}$

change .

Practice Problem

A bank branch located in a commercial district of a city has developed an improved process for serving customers during the noon-to-1:00 p.m. lunch period. The waiting time, in minutes (defined as the time the customer enters the line to when he or she reaches the teller window), of a sample of 15 customers during this hour is recorded over a period of one week.

4.21 5.55 3.02 5.13 4.77 2.34 3.54 3.20 4.50 6.10 0.38

5.12 6.46 6.19 3.79

a. Compute the mean, median, first quartile, and third quartile.

b. Compute the variance, standard deviation, range, interquartile range,. Are there any outliers? Explain.

c. Are the data skewed? If so, how?

d. As a customer walks into the branch office during the lunch hour, she asks the branch manager how long she can expect to wait. The branch manager replies, Almost certainly, less than five minutes. On the basis of the results of (a) through (c), evaluate the accuracy of this statement.

Practice Problems:

Q.2 The data given below is the total fat, in grams per serving, for a sample of 20 chicken sandwiches from fast-food chains.

7 8 4 5 16 20 20 24 19 30 23 30 25 19 29 29 30 30 40 56

a. Compute the mean, median, first quartile, and third quartile.

b. Compute the variance, standard deviation, range, interquartile range, Are there any outliers? Explain.

c. Are the data skewed? If so, how?

d. Based on the results of (a) through (c), what conclusions can you reach concerning the total fat of chicken sandwiches?

Practice Problem:

Q.3 The following data represent the battery life (in shots) for three pixel digital cameras:

300 180 85 170 380 460 260 35 380 120
110 240

- ➢ List the Five-point summary.

Q.4 For the data set below:

82	45	64	80	82	74	79	80	80	78	80	80	48	73	80	79	81	70	78	73

a. Obtain and interpret the quartiles.
b. Determine and interpret the interquartile range.
c. Find and interpret the five-number(point) summary.
d. Identify potential outliers, if any.
e. Construct and interpret a boxplot.

Practice Problem:

Q.5 A bank branch located in a commercial place of a city has developed an improved process for serving customers during the noon-to-1:00 p.m. lunch period. The waiting time, in minutes (defined as the time the customer enters the line to when he or she reaches the teller window), of a sample of 15 customers during this hour is recorded over a period of one week. The results are: 4.21, 5.55, 3.02, 5.13, 4.77, 2.34, 3.54, 3.20, 4.50, 6.10, 0.38, 5.12, 6.46, 6.19, 3.79.

Another branch, located in a residential area, is also concerned with the noon-to-1 p.m. lunch hour. The waiting time, in of a sample of 15 customers during this hour is recorded over a period of one week. The results are listed below: 9.66, 5.90, 8.02, 5.79, 8.73, 3.82, 8.01, 8.35, 10.49, 6.68, 5.64, 4.08, 6.17, 9.91, 5.47.

a. List the five-number summaries of the waiting times at the two bank branches.

b. Construct box-and-whisker plots and describe the shape of the distribution of each for the two bank branches.

c. What similarities and differences are there in the distributions of the waiting time at the two bank branches?

Thank You!