BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Overview of the course
& Basic Probability & Statistics (CS -1)

(Session 1: 20th /21st May,2023)

Overview of the course

- M 1 : Basic Probability & Statistics
- M 2 : Conditional Probability & Bayes’ theorem
- M 3 : Probability Distributions
- M 4 : Hypothesis Testing
- M 5 : Prediction & Forecasting
- M 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

No	Name	Type	Weight
EC-1(a)	Quizzes – 1 ,2 & 3 (Best two will be considered)	Online	10%
EC-1(b)	Assignments - 2	Online	20%
EC-2	Mid-Semester Test	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, Introduction to probability	T1 & T2

“Statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write”

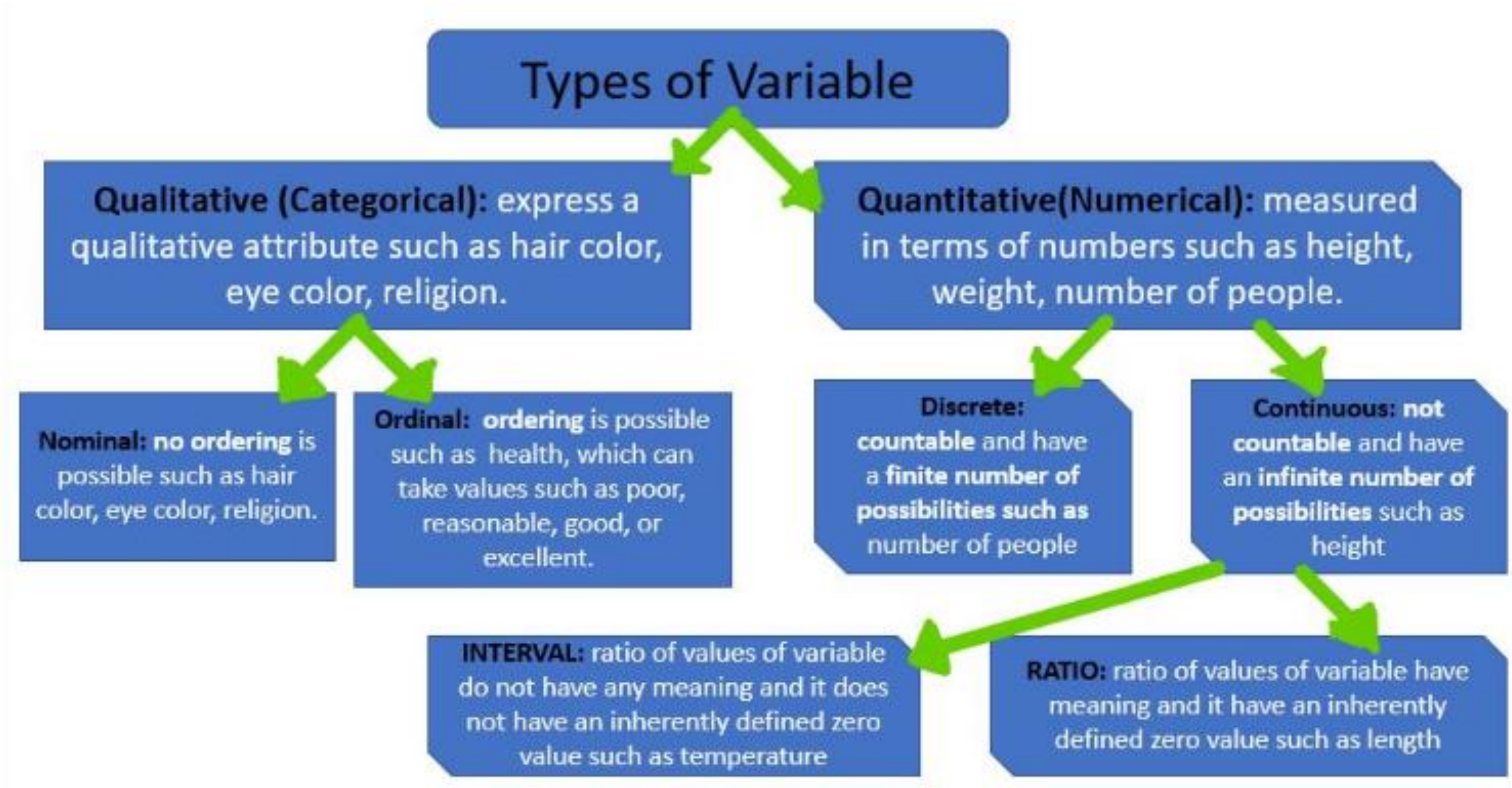
H G Wells

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.

In other words, Statistics can also be defined as numerical data with a view to analyse it.


Types of Variable
Qualitative (Categorical): express a qualitative attribute such as hair color, eye color, religion.
Quantitative(Numerical): measured in terms of numbers such as height, weight, number of people.
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.
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

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

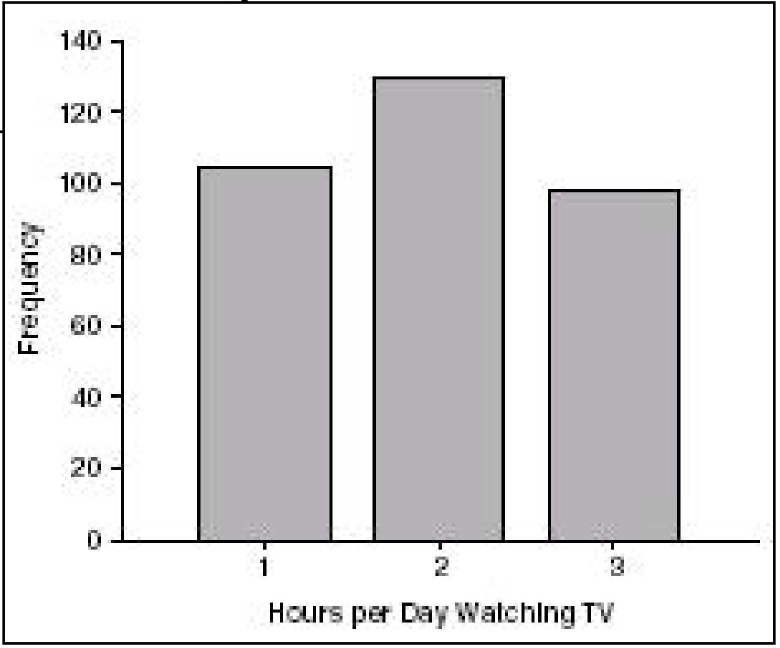
$$\bar{Y} = \frac{\sum f\,Y}{N}$$

Where f Y = a score multiplied by its frequency

Mean: Grouped Scores

Hours Spent Watching TV	Frequency (f)	fY	Percentage	C%
1	104	104	31.3	31.3
2	130	260	39.2	70.5
3	98	294	29.5	100.0
Total	332	658	100.0	

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



Data of Children watching TV in Bengaluru

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.
- 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.

Mean

When to Use the Mean

- Sampling stability is desired.
- Other measures are to be computed such as standard deviation, coefficient of variation and skewness

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:

- Number of Votes for Candidates for Lok Sabha MP. The mode, in this case, gives you the "central" response of the voters: the most popular candidate.

  - Candidate A – 11,769 votes
  - Candidate B – 39,443 votes
  - Candidate C – 78,331 votes

The Mode: "Candidate C"

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 “typical” value is desired.
  - When the data set is measured on a nominal scale

The Median

- The score that divides the distribution into two equal parts, so that half the cases are above it and half below it.

- 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.

Measures of central tendency

- The mean

$$\frac{x_1 + x_2 + x_3 + \dots x_n}{n} \quad \mu$$

$n-1 \quad \bar{x}$

Drawback?

- the median

10, 15, 20, 25, 26

10, 15, 20, 25, 28, 32

Average of there

- the mode

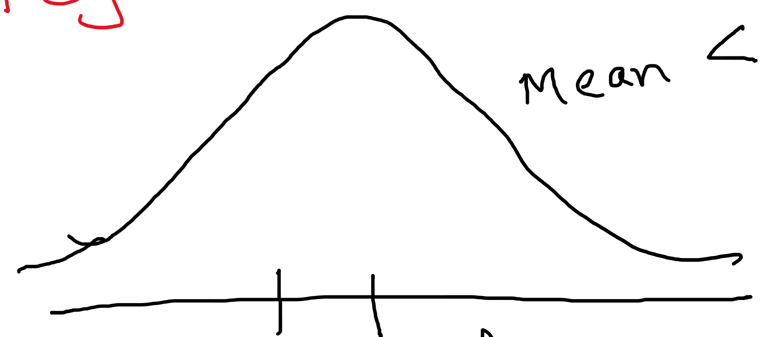
2, 5, 5, 2, 3, 2, 2, 2

5, 2, 5, 5, 2, 3, 2, 2, 2, 5, 5

Data : Symmetrical and Asymmetrical

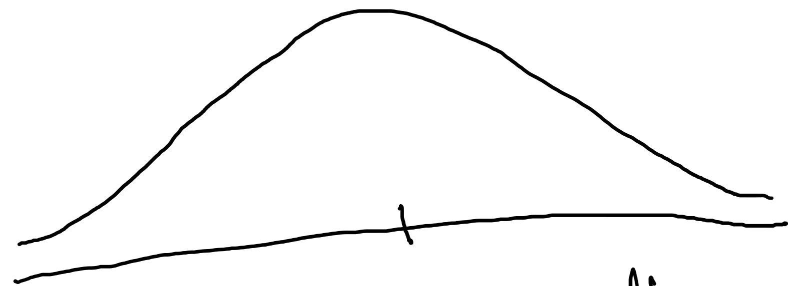
Negative Skew

Mean < Median

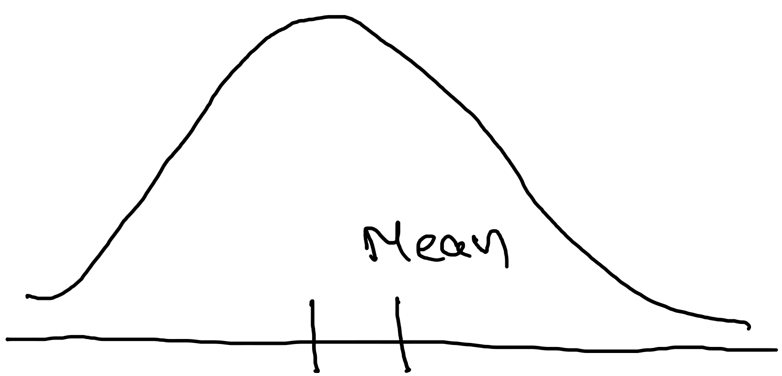


Mean Median

Symmetric



Mean = Median



Median

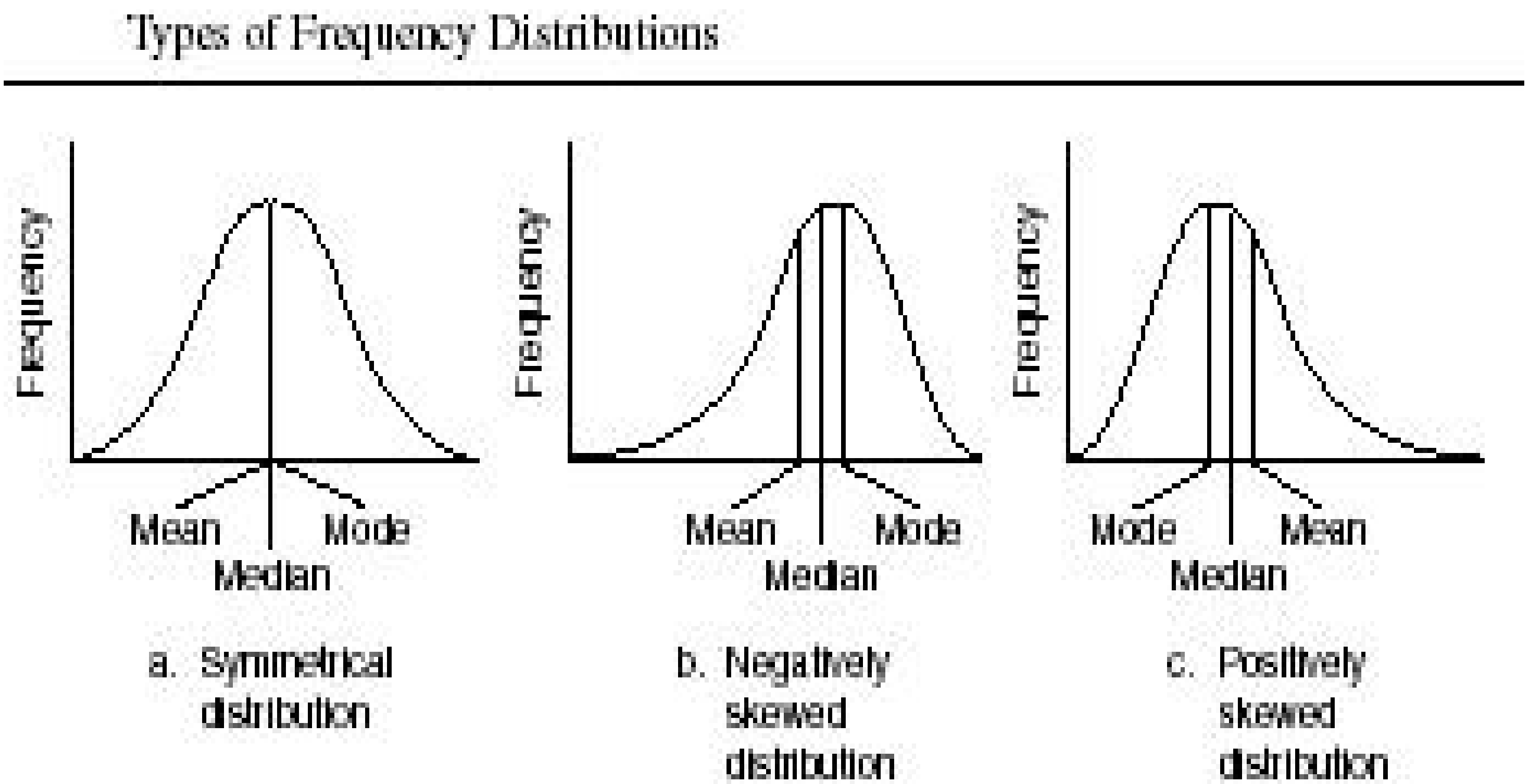
Positive Skew

Mean > Median

Shape of the distribution of data

- Symmetrical : Mean is equal to median
- Skewed
  - Negatively : mean < median
  - Positively : mean > median
- Bimodal : has two distinct modes
- Multi-modal : has more than 2 distinct modes

Distribution Shape



Sl. No.	X_1	X_2
1	2	1
2	8	15
3	5	5
4	3	5
5	7	6
6	8	3
7	5	5
8	2	2
9	5	3
Total	45	45

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	X_2
1	2	1
2	8	15
3	5	5
4	3	5
5	7	6
6	8	3
7	5	5
8	2	2
9	5	3
Total	45	45

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

Do we need any other measure?

Answer: Yes

Measures of variability

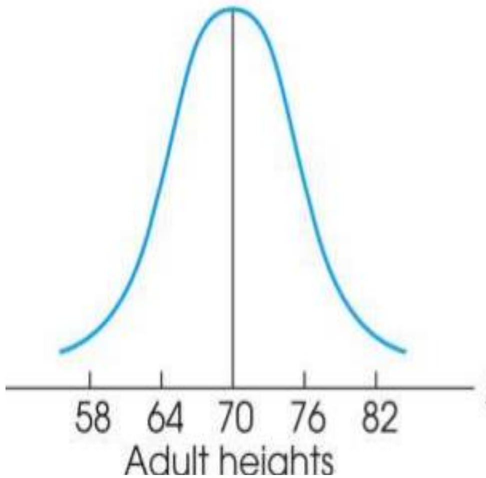
Three Measures of Variability:

- The Range
- The Variance
- The Standard Deviations

Measure of Variability

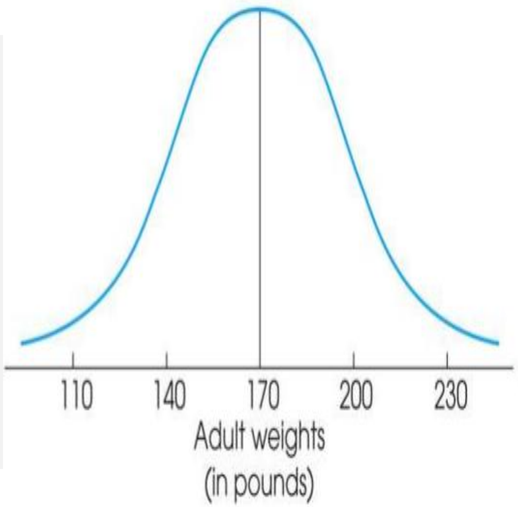
Variability can be defined several ways:

- A quantitative distance measure based on the differences between scores
- Describes distance of the spread of scores or distance of a score from the mean



Purposes of Measure of Variability:

- Describe the distribution
- Measure how well an individual score represents the distribution



The Three Measures

Three Measures of Variability:

- The Range
- The Variance
- The Standard Deviations

The Ranges

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

Range = URL for X_{max} - LRL for X_{min}

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

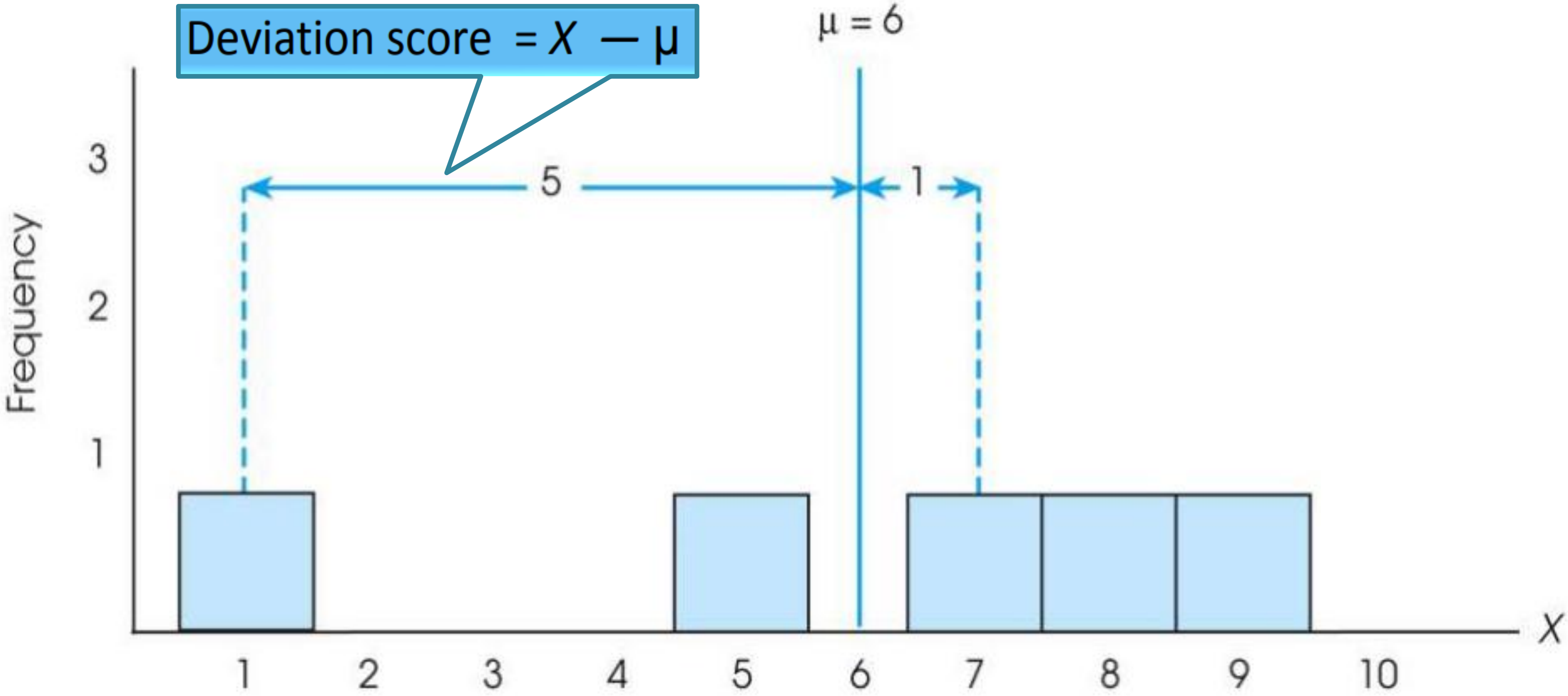
Example: For a set of scores: 7, 2, 7, 6, 5, 6, 2

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



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

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

The Variance

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

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

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

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
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

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{45}{5} = 5$$

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

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

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

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{45}{5} = 5$$

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

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

Standard Deviation and Variance for a Sample

- Goal of inferential statistics:
  - Draw general conclusions about population
  -
  - Based on limited information from a sample

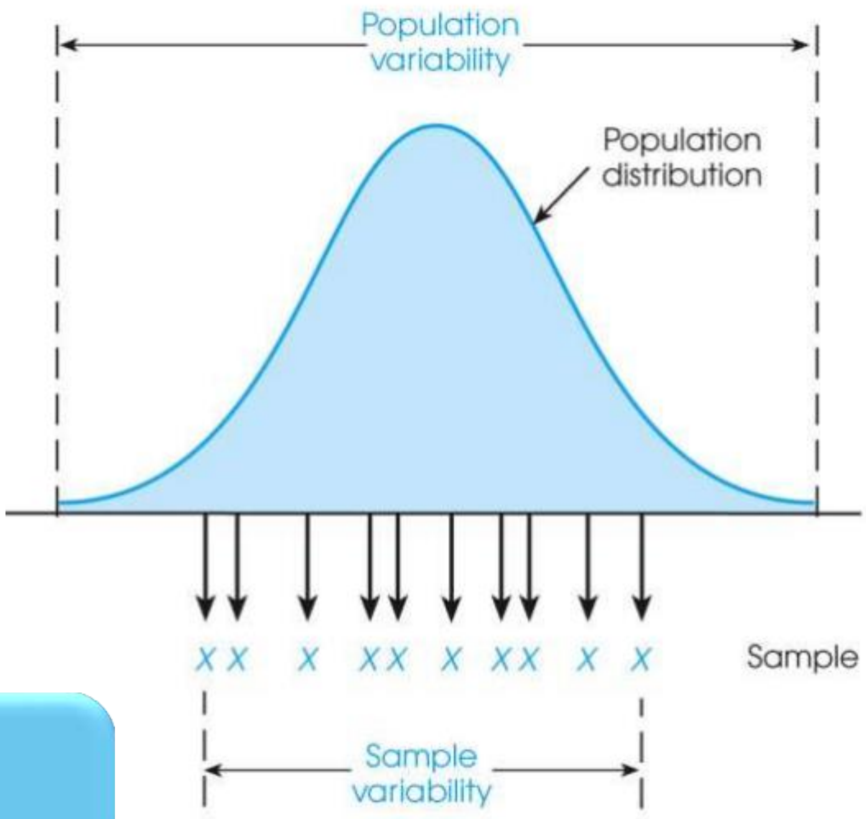
- Samples differ from the population
  - Samples have less variability
  - Computing the Variance and Standard Deviation in the same way as for a population would give a biased estimate of the population values

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

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

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



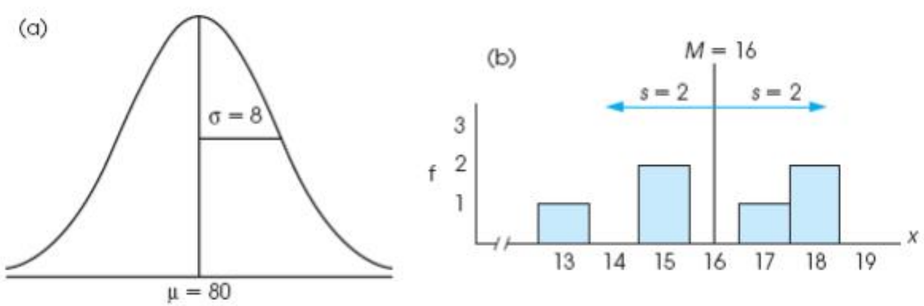
Population of Adult Heights

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

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**.

Interquartile range (IQR)

- It is measure of Variation
- Also Known as Midspread : Spread in the Middle 50%
- Difference Between Third & First Quartiles:
- Not Affected by Extreme Values

Interquartile Range = IQR = $Q_3 - Q_1$

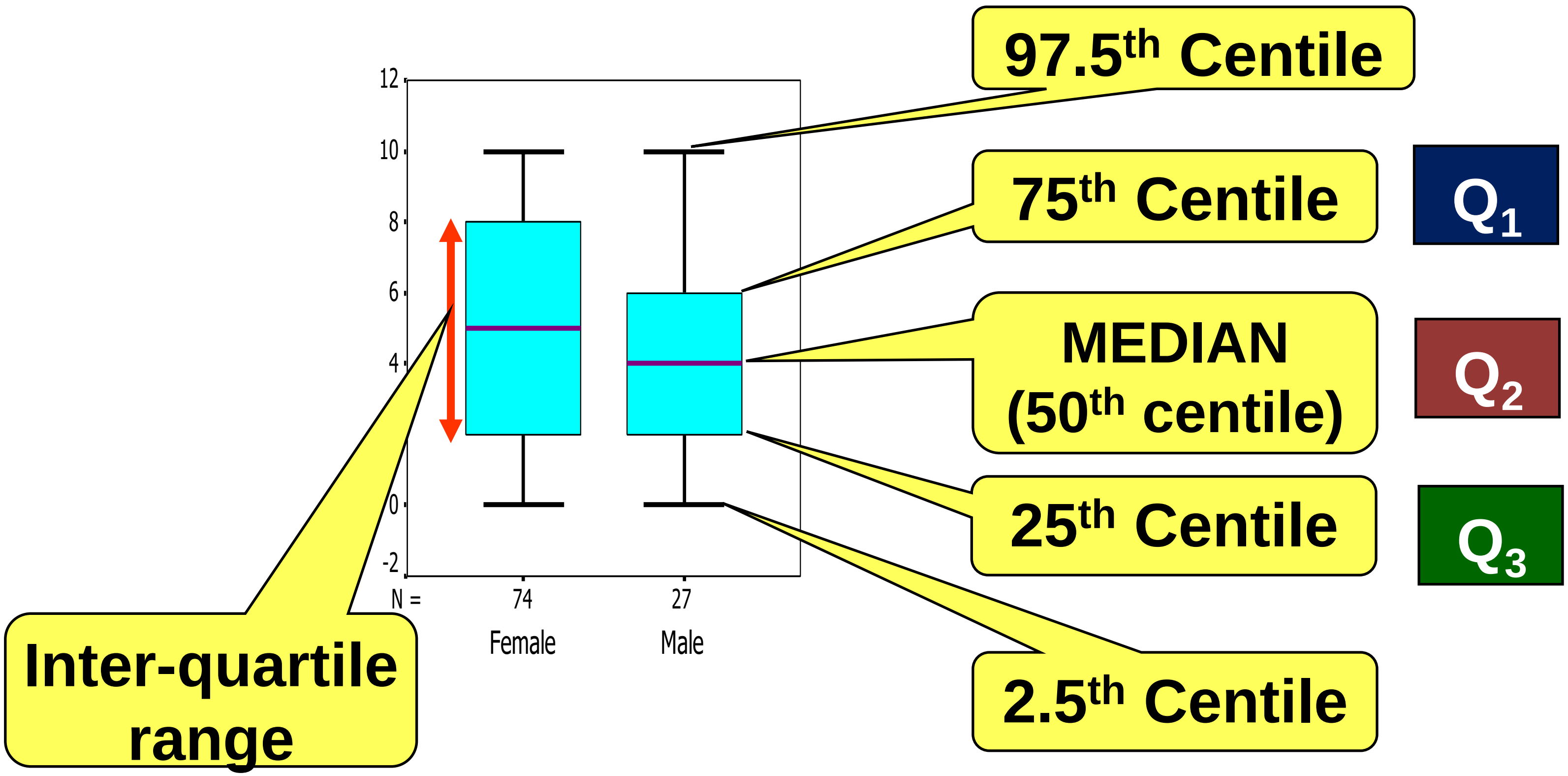
Data in Ordered Array: 11 12 13 16 16 17 17 18 21

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

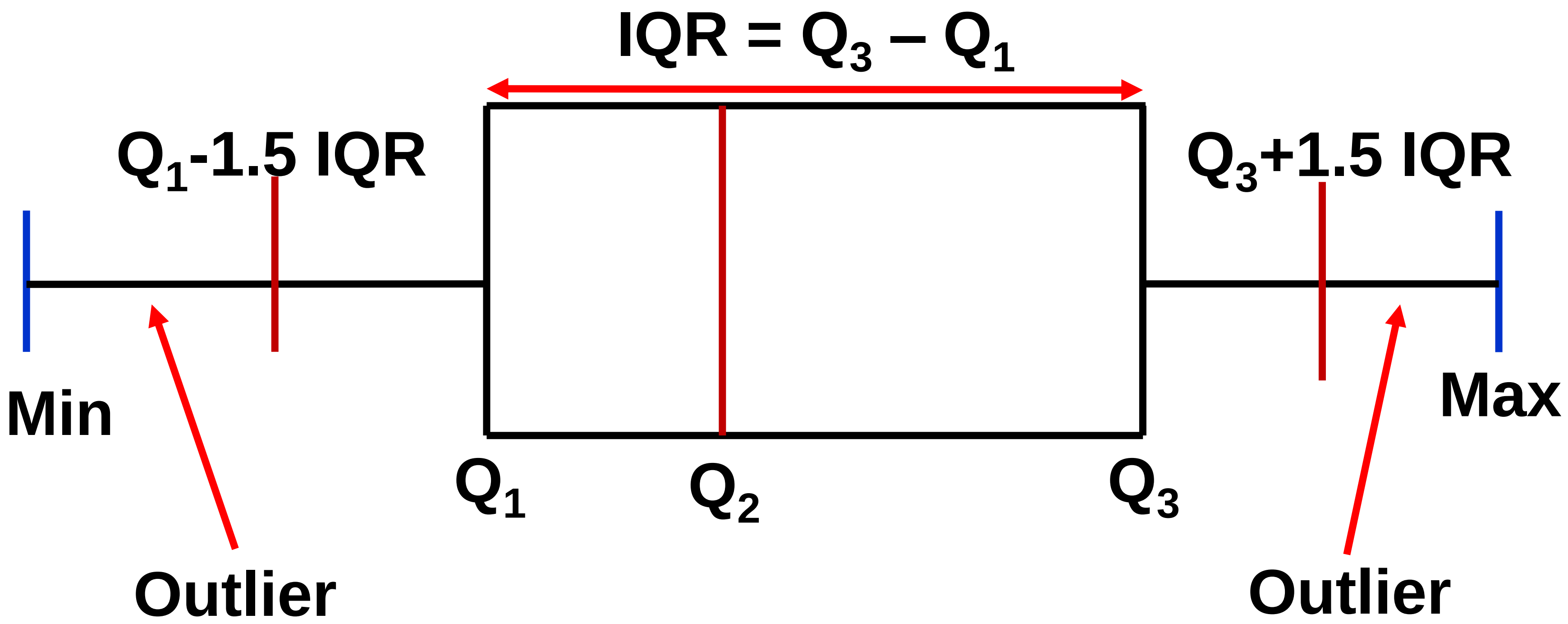
$$\text{Position of } Q_3 = \frac{3\bullet(9+1)}{4} = 7.50, \qquad Q_3 = 17.5$$

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

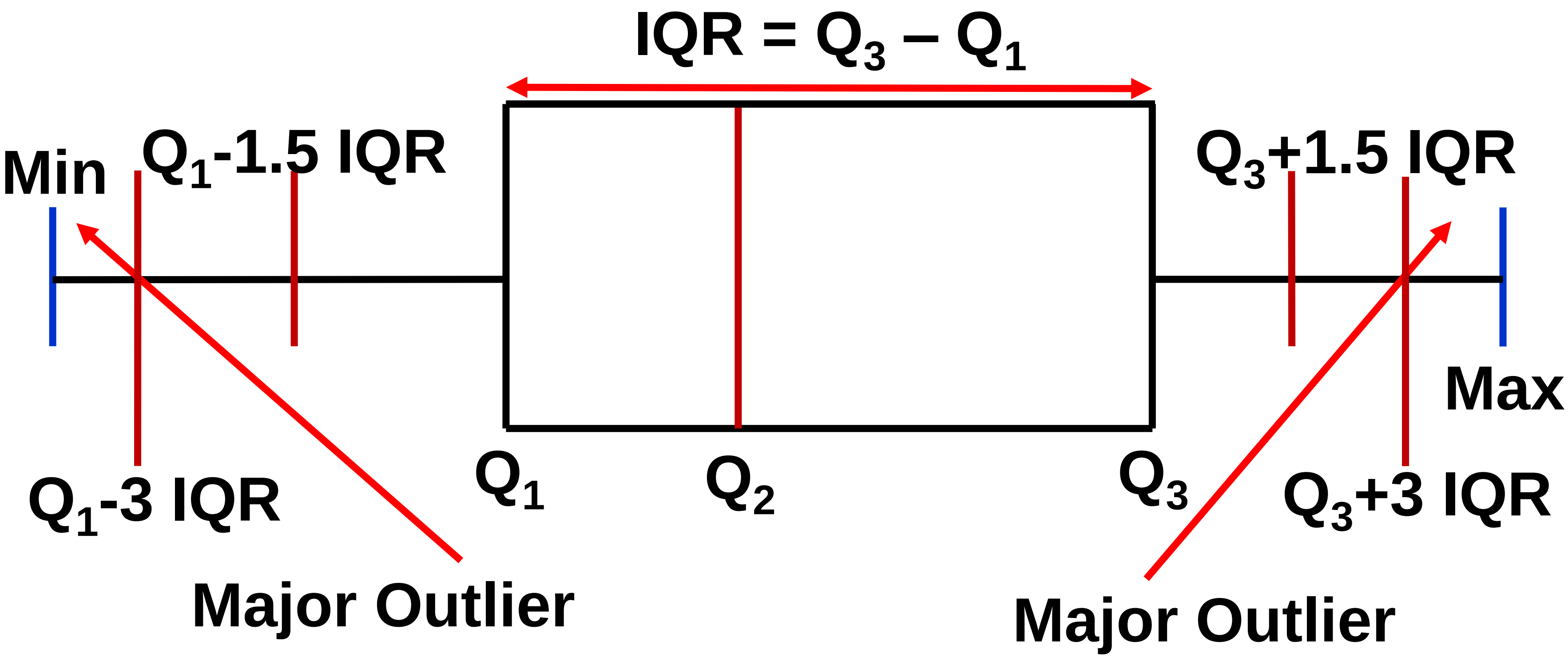
Box-and-Whisker plot



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.

HW problem :

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.

HW problem :

Human measurements provide a rich area of application for statistical methods. The article "A Longitudinal Study of the Development of Elementary School Children's Private Speech" (*Merrill-Palmer Q.*, 1990: 443–463) reported on a study of children talking to themselves (private speech). It was thought that private speech would be related to IQ, because IQ is supposed to measure mental maturity, and it was known that private speech decreases as students progress through the primary grades. The study included 33 students whose first-grade IQ scores are given here:

82	96	99	102	103	103	106	107	108	108	108	108
109	110	110	111	113	113	113	113	115	115	118	118
119	121	122	122	127	132	136	140	146			

Describe the data and comment on any interesting features.

Introduction to probability

Random Experiment :

❖ The term "**random experiment**" is used to describe any action whose outcome is not known in advance. Here are some examples of experiments dealing with statistical data:

- Tossing a coin
- Counting how many times a certain word or a combination of words appears in the text of the “King Lear” or in a text of Confucius.
- Counting occurrences of a certain combination of amino acids in a protein database.
- Pulling a card from the deck.

the experiment.

Sample spaces and events

❖ If the experiment is to throw a die and record the outcome, then sample space is S = { 1,2,3,4,5,6}

Event : An event is a subset of sample space of the random experiment.

Definition of probability

Classical approach :

CLASSICAL PROBABILITY $$\text{Probability of an event} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Empirical approach :

Empirical or **relative frequency** is the second type of objective probability. It is based on the number of times an event occurs as a proportion of a known number of trials.

EMPIRICAL PROBABILITY The probability of an event happening is the fraction of the time similar events happened in the past.

In terms of a formula:

$$\text{Empirical probability} = \frac{\text{Number of times the event occurs}}{\text{Total number of observations}}$$

The empirical approach to probability is based on what is called the law of large numbers. The key to establishing probabilities empirically is that more observations will provide a more accurate estimate of the probability.

LAW OF LARGE NUMBERS Over a large number of trials, the empirical probability of an event will approach its true probability.

innovate achieve lead

Axiomatic approach :

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:

If S is the sample space and E is any event in a random experiment,

(1) $P(S) = 1$

(2) $0 \leq P(E) \leq 1$

(3) For two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Thank You