Allama Iqbal Open University Islamabad

Semester: Spring 2025

Course: Introduction to Statistics (4485)

NAME	
ID	
PROGRAM	
SEMESTER	SPRING 2025

ASSIGNMENT No. 1

Q. 1 (a) Define statistics. Discuss, giving examples, the importance of the study of statistics and show how it can help the extension of scientific knowledge.

(b)Describe the methods which can be used in the collection of statistical data ,stating the advantages and disadvantages of each method.

Definition of Statistics: Statistics is the science and practice of developing human knowledge through the use of empirical data. It involves the collection, organization, analysis, interpretation, and presentation of data. Statistics is broadly divided into two main branches:

Descriptive Statistics: This involves summarizing and describing the main features of a collection of data. Techniques include calculating measures like mean, median, mode, standard deviation, and creating graphical representations like histograms, bar charts, and pie charts. Inferential **Statistics:** This involves drawing conclusions, making predictions, or inferring properties about a larger population based on data collected from a smaller sample. Techniques include hypothesis testing, confidence intervals, regression analysis, and analysis of variance (ANOVA). Importance and Examples: The study of statistics is crucial in virtually every field because it provides the tools to deal with variability and uncertainty, enabling informed decision-making.

Business and Economics: Businesses use statistics for market research (understanding customer preferences), quality control (ensuring product standards), financial analysis (stock market trends, risk assessment), and sales forecasting. Economists use it to analyze economic trends, forecast GDP, unemployment rates, and inflation, and to evaluate the impact of economic policies. Healthcare and Medicine: Statistics is vital for clinical trials (testing the effectiveness and safety of new drugs or treatments), epidemiology (studying the patterns, causes, and effects of health and disease conditions), and health services research (improving healthcare delivery). Social Sciences: Researchers use statistics to analyze survey data (opinion polls, social attitudes), study demographic trends (population growth, migration), and evaluate the effectiveness of social programs. Engineering and Technology: Statistics helps in reliability analysis (predicting product lifespan), process control (monitoring and improving manufacturing processes), and designing experiments to optimize performance. Government: Governments rely heavily on statistics for policy-making, resource allocation, census data collection, and monitoring national well-being. Everyday Life: Understanding basic statistics helps individuals interpret news reports (e.g., understanding polls, health risks reported in studies), manage personal finances, and make better decisions based on data encountered daily. Extension of Scientific Knowledge: Statistics is fundamental to the scientific method and the extension of scientific knowledge:

Designing Experiments: Statistical principles guide the design of experiments (e.g., sample size determination, randomization, control groups) to ensure that data collected is relevant and can lead to valid conclusions. Objective Data Analysis: Statistics provides objective methods to analyze data, reducing the influence of researcher bias. It allows scientists to identify patterns, relationships, and trends that might not be obvious. Hypothesis Testing: It offers a formal framework for testing scientific hypotheses. Researchers can determine whether observed results are likely due to chance or represent a real effect, quantifying the level of certainty (e.g., using p-values). Quantifying Uncertainty: Scientific findings are rarely absolute. Statistics allows researchers to quantify the uncertainty associated with their results using confidence intervals and margins of error, providing a more realistic understanding of the findings. Model Building: Statistical models (like regression) help scientists understand complex relationships between variables and make predictions, contributing to theoretical development. Reproducibility: Proper statistical analysis and reporting are crucial for the reproducibility of scientific research, allowing other scientists to verify findings. Q. 1 (b) Describe the methods which can be used in the collection of statistical data, stating the advantages and disadvantages of each method.

B Methods for collecting statistical data can be broadly categorized into primary and secondary data collection.

Primary Data Collection Methods: (Collecting new data)

Observation: Researcher observes and records behavior, events, or characteristics in their natural setting or a controlled environment. Advantages: Provides direct information, captures behavior as it occurs, useful in natural settings. Disadvantages: Prone to observer bias, can be time-consuming, doesn't reveal attitudes or opinions, ethical concerns if observation is covert. Interviews: Collecting data through direct verbal interaction between interviewer and respondent. Can be structured (fixed questions), semi-structured, or unstructured. Personal Interviews (Face-to-face): Advantages: High response rate, allows clarification of questions, enables collection of detailed/complex information, allows use of visual aids. Disadvantages: Expensive (time, travel), potential for interviewer bias, less anonymity for respondent. Telephone Interviews: Advantages: Less expensive and quicker than personal interviews, wider geographical reach possible. Disadvantages: Lower response rate than personal interviews, limited interview length, cannot use visual aids, excludes households without phones. Questionnaires: Respondents answer a set of written questions. Mail Questionnaires: Advantages: Relatively inexpensive, can reach a wide area, respondent anonymity is high. Disadvantages: Low response rates, no opportunity to clarify questions, requires respondent literacy, slow data collection. Online Questionnaires/Surveys: Advantages: Very low cost, fast data collection, wide reach, data can be automatically captured. Disadvantages: Requires internet access, potential for sampling bias (not everyone is online), survey fatigue, concerns about data security/authenticity. Experiments: Manipulating one or more independent variables to observe their effect on a dependent variable, while controlling other factors. Advantages: Allows establishing cause-and-effect relationships, high level of control over variables. Disadvantages: Can be artificial (may not reflect real-world conditions), ethical limitations, can be complex and expensive to conduct. Secondary Data Collection Methods: (Using existing data)

Using Published Sources/Databases: Collecting data from government publications (e.g., census reports, economic data), academic journals, organizational records, online databases, books, etc. Advantages: Quick and relatively inexpensive to obtain, often covers large populations or long time periods. Disadvantages: Data may not exactly fit the research needs (different definitions, units), data quality/accuracy may be unknown or questionable, data can be outdated.

Q. 2 (a)Explain what is meant by classification. What are its basic principles?

(b)Arrange the data given below in an array and construct a frequency distribution, using a class interval of 5.00. Indicate the class boundaries and class limits.

79.4	**71.6**	**95.5**	**73.0**	**74.2**	**81.8**	**90.6**	**55.9**
75.2	**81.9**	**68.9**	**74.2**	**80.7**	**65.7**	**67.6**	**82.9**
88.1	**77.8**	**69.4**	**83.2**	**82.7**	**73.8**	**64.2**	**63.9**
68.3	**48.6**	**83.5**	**70.8**	**72.1**	**71.6**	**59.4**	**77.6**

- **Explanation of Classification:** Classification, in statistics, is the process of arranging data into groups or classes based on their common characteristics or attributes. It involves systematically organizing raw data into a more compact, understandable, and manageable form. The primary goal is to condense the data, highlight similarities and differences, facilitate comparisons, identify underlying patterns, and prepare the data for further statistical analysis. For example, classifying a population by age group, income level, or educational attainment.

- **Basic Principles (Characteristics of a Good Classification):**
    1. **Exhaustive:** Every single item in the dataset must belong to at least one class. No item should be left out.
    2. **Mutually Exclusive:** Each item in the dataset must belong to only one class. The classes should not overlap.
    3. **Unambiguity/Clarity:** The definition of each class must be clear, precise, and unambiguous. There should be no confusion about where any given item belongs.
    4. **Suitability:** The basis chosen for classification should align with the objective of the study or inquiry.
    5. **Stability:** The basis of classification should remain consistent throughout the analysis. Changing the criteria mid-way invalidates the classification.
    6. **Homogeneity:** Items within any given class should be as similar (homogeneous) as possible with respect to the characteristic being used for classification.
    7. **Flexibility (Elasticity):** The classification system should be flexible enough to accommodate new data or minor changes without altering the entire structure significantly.
    8. **Appropriate Number of Classes:** The number of classes should generally not be too large or too small (often suggested between 5 and 15 classes) to effectively summarize the data without losing too much detail.

Q. 2 (b) Arrange the data given below in an array and construct a frequency distribution, using a class interval of 5.00. Indicate the class boundaries and class limits.

Data: 79.4, 71.6, 95.5, 73.0, 74.2, 81.8, 90.6, 55.9, 75.2, 81.9, 68.9, 74.2, 80.7, 65.7, 67.6, 82.9, 88.1, 77.8, 69.4, 83.2, 82.7, 73.8, 64.2, 63.9, 68.3, 48.6, 83.5, 70.8, 72.1, 71.6, 59.4, 77.6 Total number of observations (N) = 32

Step 1: Arrange the data in an array (ascending order): 48.6, 55.9, 59.4, 63.9, 64.2, 65.7, 67.6, 68.3, 68.9, 69.4, 70.8, 71.6, 71.6, 72.1, 73.0, 73.8, 74.2, 74.2, 75.2, 77.6, 77.8, 79.4, 80.7, 81.8, 81.9, 82.7, 82.9, 83.2, 83.5, 88.1, 90.6, 95.5

Step 2: Determine the Range: Range = Maximum Value - Minimum Value = 95.5 - 48.6 = 46.9

Step 3: Determine the Number of Classes: Class Interval (width), h = 5.00 (given) Approximate Number of Classes = Range / h = 46.9 / 5.00 ≈ 9.38. We will likely need around 10 classes.

Step 4: Define Class Limits and Class Boundaries: We need classes that cover the range from 48.6 to 95.5 with a width of 5.00. Let's start the first class slightly below the minimum value, for instance, at 48.0. Since the data has one decimal place, the precision is 0.1. The gap between the upper limit of one class and the lower limit of the next should be 0.1. The boundaries will be midway between these limits.

Let's try limits:

- 48.0 - 52.9 (Width = 52.9 - 48.0 = 4.9 - This doesn't seem right for interval 5.00)

Let's define classes such that the lower limit of a class plus the interval gives the lower limit of the next class. Start at 48.0:

- Class 1: 48.0 - 52.9 (Includes values ≥48.0 and <53.0)
- Class 2: 53.0 - 57.9
- Class 3: 58.0 - 62.9
- Class 4: 63.0 - 67.9
- Class 5: 68.0 - 72.9
- Class 6: 73.0 - 77.9
- Class 7: 78.0 - 82.9
- Class 8: 83.0 - 87.9
- Class 9: 88.0 - 92.9
- Class 10: 93.0 - 97.9

Now let's find the boundaries. The adjustment factor is half the gap between classes: (53.0 - 52.9) / 2 = 0.1 / 2 = 0.05. Subtract 0.05 from lower limits and add 0.05 to upper limits.

- Boundaries for Class 1: 48.0 - 0.05 = 47.95 and 52.9 + 0.05 = 52.95. Width = 52.95 - 47.95 = 5.00. This works.

Step 5: Tally the Data and Construct the Frequency Distribution:

Class Limits	Class Boundaries	Tally	Frequency (f)
48.0 - 52.9	47.95 - 52.95		
53.0 - 57.9	52.95 - 57.95		
58.0 - 62.9	57.95 - 62.95		
63.0 - 67.9	62.95 - 67.95		
68.0 - 72.9	67.95 - 72.95		
73.0 - 77.9	72.95 - 77.95		
78.0 - 82.9	77.95 - 82.95		
83.0 - 87.9	82.95 - 87.95		
88.0 - 92.9	87.95 - 92.95		
93.0 - 97.9	92.95 - 97.95		
Total			**N = 32**

Summary:

- **Array:** 48.6, 55.9, 59.4, 63.9, 64.2, 65.7, 67.6, 68.3, 68.9, 69.4, 70.8, 71.6, 71.6, 72.1, 73.0, 73.8, 74.2, 74.2, 75.2, 77.6, 77.8, 79.4, 80.7, 81.8, 81.9, 82.7, 82.9, 83.2, 83.5, 88.1, 90.6, 95.5
- **Frequency Distribution Table:** (As shown above)

- **Class Limits:** These are the stated minimum and maximum values for each class (e.g., 48.0 - 52.9, 53.0 - 57.9, etc.).
- **Class Boundaries:** These are the true dividing points between classes, eliminating the gap between consecutive class limits (e.g., 47.95 - 52.95, 52.95 - 57.95, etc.). They ensure continuity for calculation purposes.

Q. 3 (a) Define the arithmetic mean. What are its advantages and limitations in the analysis of data? Give various methods of calculating the arithmetic mean.

(b)The following table represents the ages of participants in a marathon race:

Age Group (years)	Number of Participants
20-29	15
30-39	25
40-49	30
50-59	20
60-69	10

Calculate the mean and median for the ages of the participants in the marathon race. Definition: The arithmetic mean (often simply called the "mean" or "average") is a measure of central tendency. It is calculated by summing all the values in a dataset and dividing by the total number of values.

For ungrouped data ($x_1, x_2, \ldots, x_n$):

$\bar{x}$

$n \sum_{i=1}^{n} x_i$

$n\ x_1 + x_2 + \ldots + x_n$

For grouped data (frequency distribution):

$\bar{x}$

$\sum_{i=1}^{k} f_i$

$\sum_{i=1}^{k} f_i x_i$

$N \sum fx$ where f_i is the frequency of the i-th class, x_i is the midpoint of the i-th class, k is the number of classes, and $N=\sum f_i$ is the total number of observations. Advantages:

Easy to Understand and Calculate: The concept is simple and the calculation is straightforward. Based on All Observations: It takes every value in the dataset into account. Unique: A given dataset has only one arithmetic mean. Suitable for Further Algebraic Treatment: It can be readily used in other statistical calculations (like standard deviation). Certainty: It's a rigidly defined measure. Limitations:

Affected by Extreme Values (Outliers): Very large or very small values can significantly distort the mean, making it unrepresentative of the typical value. Cannot be Calculated for Open-Ended Classes: If a frequency distribution has open-ended classes (e.g., "less than 20" or "60 and over"), the midpoint cannot be determined, and thus the mean cannot be calculated accurately using the standard grouped data formula. May Not Exist in the Data: The calculated mean value might not actually be one of the values present in the dataset (especially for discrete data). Cannot be Determined Graphically: Unlike the median or mode, the mean cannot be easily estimated from graphical representations like histograms. Misleading for Skewed Distributions: In highly skewed distributions, the mean is pulled towards the tail and may not be the best measure of central tendency compared to the median. Methods of Calculation (for Grouped Data):

Direct Method: Uses the formula

x ˉ

N ∑fx , where x represents the class midpoints. This is straightforward but can involve large numbers if data values or frequencies are high. Short-cut Method (Assumed Mean Method): Uses the formula x ˉ =A+ N ∑fd , where: A is an assumed mean (usually the midpoint of a class near the center of the distribution). d=x−A is the deviation of each class midpoint (x) from the assumed mean (A). This method simplifies calculations by working with smaller deviation values. Step-Deviation Method: Used when class intervals are equal. It's a further simplification of the short-cut method. The formula is x ˉ =A+(N ∑fu)×h, where: A is the assumed mean. h is the common class interval width. u= h x−A

h d is the step deviation. This method involves the smallest numerical values, making manual calculations easiest. Q. 3 (b) The following table represents the ages of participants in a marathon race: [Table provided] Calculate the mean and median for the ages of the participants in the marathon race.

Data Table:

Age Group (years) Number of Participants (f) 20-29 15 30-39 25 40-49 30 50-59 20 60-69 10 Total N = 100

Calculation of the Mean (using Direct Method):

Find Class Midpoints (x): 20-29: (20 + 29) / 2 = 24.5 30-39: (30 + 39) / 2 = 34.5 40-49: (40 + 49) / 2 = 44.5 50-59: (50 + 59) / 2 = 54.5 60-69: (60 + 69) / 2 = 64.5 Calculate fx: 15 * 24.5 = 367.5 25 * 34.5 = 862.5 30 * 44.5 = 1335.0 20 * 54.5 = 1090.0 10 * 64.5 = 645.0 Calculate Σfx: Σfx = 367.5 + 862.5 + 1335.0 + 1090.0 + 645.0 = 4300.0 Calculate Mean (x̄): x ˉ

N ∑fx

100 4300.0 =43 The mean age of the participants is 43 years.

Calculation of the Median:

Find the Median Position: N / 2 = 100 / 2 = 50. The median is the value of the 50th participant when arranged in order.

Determine Class Boundaries and Cumulative Frequency (CF): Assuming integer ages, the boundaries are: 19.5, 29.5, 39.5, 49.5, 59.5, 69.5. Class width (h) = 29.5 - 19.5 = 10.

Age Group	Boundaries	Frequency (f)	Cumulative Frequency (CF)
20-29	19.5 - 29.5	15	15
30-39	29.5 - 39.5	25	15 + 25 = 40
40-49	39.5 - 49.5	30	40 + 30 = 70
50-59	49.5 - 59.5	20	70 + 20 = 90
60-69	59.5 - 69.5	10	90 + 10 = 100

Identify the Median Class: The class whose cumulative frequency is the first to equal or exceed N/2 (50). This is the 40-49 class (CF = 70).

Identify Variables for Median Formula:

L = Lower boundary of the median class = 39.5 N = Total frequency = 100 CF_prev = Cumulative frequency of the class preceding the median class = 40 f = Frequency of the median class = 30 h = Class width = 10 Apply the Median Formula: Median=L+(f N/2−CF prev

)×h Median=39.5+(30 100/2−40)×10 Median=39.5+(30 50−40)×10 Median=39.5+(30 10)×10 Median=39.5+ 30 100

Median=39.5+3.333... Median≈42.83

The median age of the participants is approximately 42.83 years.

Q. 4 (a) Define the mode of a frequency distribution. How does it compare with other types of averages?

(b) A survey asked 100 people their favourite color, and the results were: 25 people chose blue, 30 chose green, 20 chose red, 10 chose yellow, 10 chose orange, and 5 chose purple. What is the mode of the data?

The **mode** of a frequency distribution is the value or category that appears most frequently in the dataset. In simpler terms, it is the observation with the highest frequency. A dataset can have one mode (unimodal), more than one mode (multimodal), or no mode if all values appear with the same frequency. The mode is particularly useful for categorical data, where calculating means or medians is not possible.

Comparison with other types of averages:

- **Mean (Arithmetic Mean):** The mean is calculated by summing all the values in a dataset and dividing by the number of observations. It represents the average value. Unlike the mode, the mean is affected by every value in the dataset, including outliers, which can skew its value. It is suitable for numerical data and assumes an interval or ratio scale of measurement. The mode does not require numerical data and is not influenced by extreme values.

- **Median:** The median is the middle value in a dataset that has been ordered from least to greatest. If there is an even number of observations, the median is typically the average of the two middle values. The median divides the dataset into two equal halves. Like the mode, the median is not affected by extreme outliers. It is suitable for numerical data and is particularly useful for skewed distributions where the mean might be misleading. The median requires the data to be ordered, while the mode only requires counting frequencies.

In summary:

- The **mode** is the most frequent value, suitable for all data types (categorical, ordinal, interval, ratio), and is not affected by outliers.
- The **mean** is the average value, suitable for interval or ratio data, and is sensitive to outliers.
- The **median** is the middle value of ordered data, suitable for ordinal, interval, or ratio data, and is not affected by outliers.

Q. 4 (b) A survey asked 100 people their favourite color, and the results were: 25 people chose blue, 30 chose green, 20 chose red, 10 chose yellow, 10 chose orange, and 5 chose purple. What is the mode of the data?

To find the mode, we need to identify the color with the highest frequency (the largest number of people who chose it).

- Blue: 25 people
- Green: 30 people
- Red: 20 people
- Yellow: 10 people
- Orange: 10 people
- Purple: 5 people

Comparing the frequencies, the highest frequency is 30, which corresponds to the color green.

Therefore, the mode of the data is **Green**.

Q. 5 (a) Define geometric mean. How does it differ from the arithmetic mean? What are its advantages and disadvantages?

(b) Calculate G.M and H.M for the following frequency distribution given below

Classes	**0-5**	**5-10**	**10-15**	**15-20**	**20-25**	**25-30**	**30-35**
Frequency	**2**	**5**	**7**	**13**	**21**	**16**	**8**

The geometric mean (G.M.) is a type of average that is calculated by multiplying all the values in a dataset and then taking the nth root of the product, where n is the number of values. Mathematically, for a set of n values x 1 ,x 2 ,...,x n , the geometric mean is given by: G.M.= n

x 1 ×x 2 ×...×x n

This formula is only applicable for positive values. For frequency distributions with class intervals, we typically use the midpoints of the classes and a modified formula involving logarithms.

Difference from the arithmetic mean:

The key difference lies in how they are calculated and what they represent.

The arithmetic mean is based on the sum of the values and is appropriate when dealing with sums or additive relationships. The geometric mean is based on the product of the values and is appropriate when dealing with products, rates of change, or multiplicative relationships (e.g., calculating average growth rates, average rates of return for investments). The geometric mean is always less than or equal to the arithmetic mean for a given set of non-negative numbers. Advantages of the Geometric Mean:

It is particularly useful for calculating average rates of change or growth over time. It is less affected by extremely large values compared to the arithmetic mean. It is appropriate when dealing with data that are multiplied together to produce a result, such as in calculating average investment returns. Disadvantages of the Geometric Mean:

It can only be calculated for positive values. If any value in the dataset is zero or negative, the geometric mean cannot be calculated using the standard formula. It is more complex to calculate than the arithmetic mean. It may not be intuitive for people who are not familiar with the concept of geometric averaging. Q. 5 (b) Calculate G.M and H.M for the following frequency distribution given below

Classes 0-5 5-10 10-15 15-20 20-25 25-30 30-35 Frequency 2 5 7 13 21 16 8

B To calculate the Geometric Mean and Harmonic Mean for a frequency distribution with class intervals, we first need to find the midpoint (x_i) of each class interval.

Classes Midpoint (x_i) Frequency (f_i) 0-5 2.5 2 5-10 7.5 5 10-15 12.5 7 15-20 17.5 13 20-25 22.5 21 25-30 27.5 16 30-35 32.5 8

Total frequency (N) = $\sum f_i = 2+5+7+13+21+16+8=72$

Calculating the Geometric Mean (G.M.)

For a frequency distribution, the G.M. is calculated using logarithms: G.M.=antilog(N $\sum(f_i \log x_i)$)

First, calculate $\log x_i$ and $f_i \log x_i$:

Midpoint (x_i) $\log x_i$ (approx) Frequency (f_i) $f_i \log x_i$ (approx) 2.5 0.3979 2 0.7958 7.5 0.8751 5 4.3755 12.5 1.0969 7 7.6783 17.5 1.2430 13 16.1590 22.5 1.3522 21 28.4062 27.5 1.4393 16 23.0288 32.5 1.5119 8 12.0952

$\sum(f_i \log x_i) = 0.7958+4.3755+7.6783+16.1590+28.4062+23.0288+12.0952=92.5388$

G.M.=antilog(72 92.5388)=antilog(1.28526)

Using a calculator, antilog(1.28526) ≈19.29

So, the Geometric Mean is approximately 19.29.

Calculating the Harmonic Mean (H.M.)

For a frequency distribution, the H.M. is calculated as: H.M.= ∑(x_i

f_i

) N

First, calculate

x i

f i

:

Midpoint (x i) Frequency (f i) x i

f i

(approx) 2.5 2 2.5 2 =0.8 7.5 5 7.5 5 =0.6667 12.5 7 12.5 7 =0.56 17.5 13 17.5 13 =0.7429 22.5 21 22.5 21 =0.9333 27.5 16 27.5 16 =0.5818 32.5 8 32.5 8 =0.2462

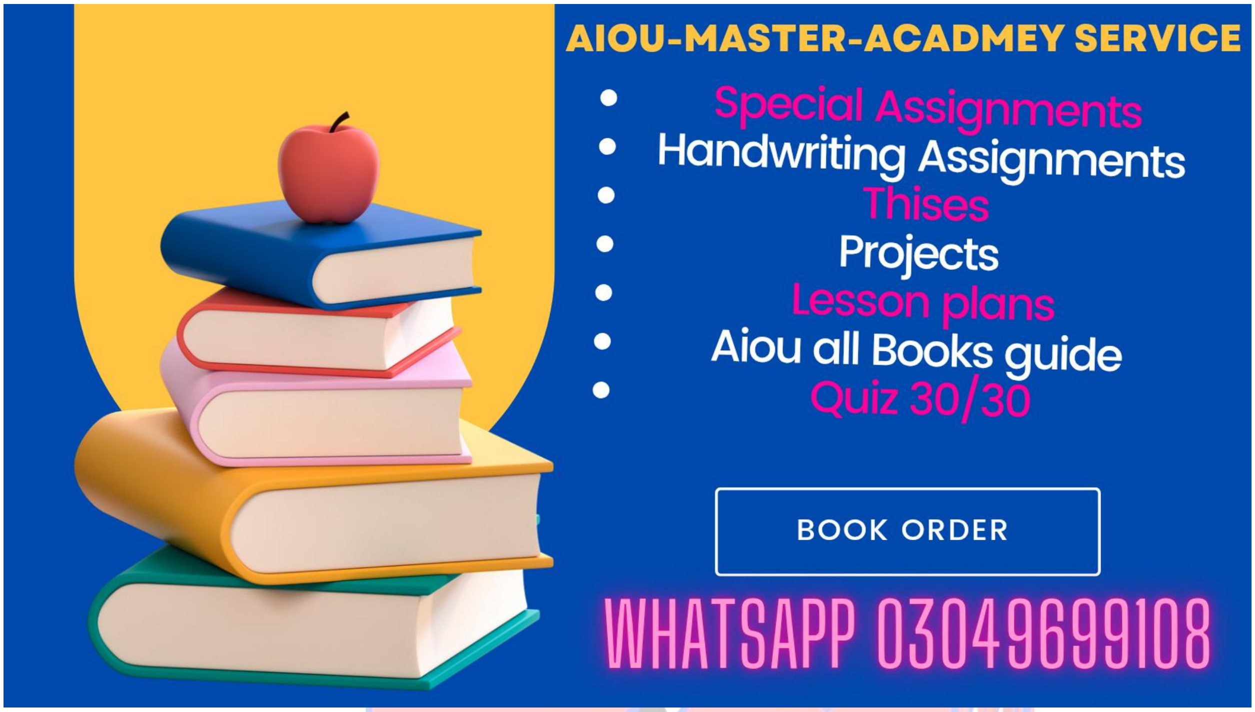
∑(x i

f i

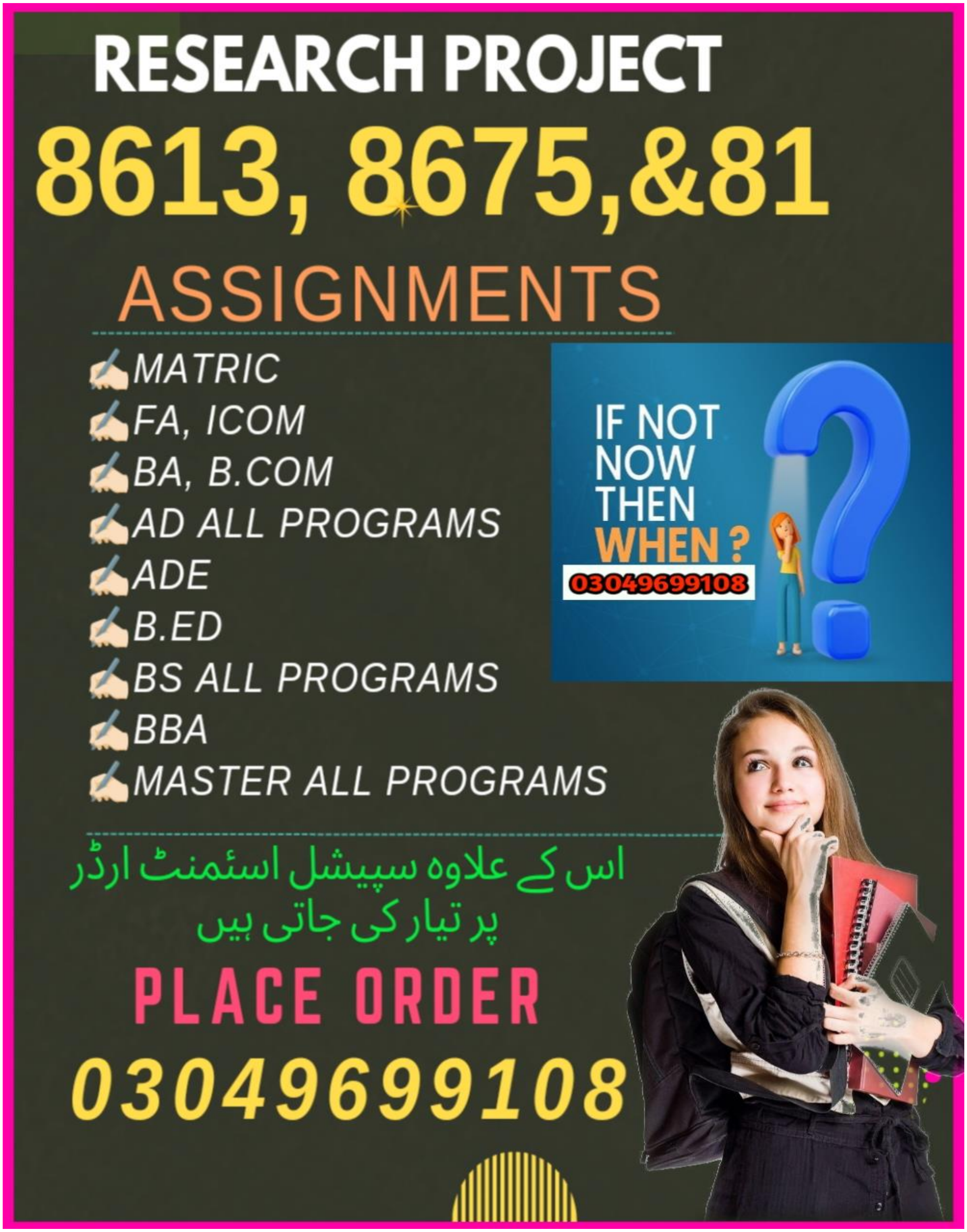
)=0.8+0.6667+0.56+0.7429+0.9333+0.5818+0.2462=4.5309

H.M.= 4.5309 72 ≈15.90

So, the Harmonic Mean is approximately 15.90.



Download free assignment PDF https://www.aioumasteracadmey.com



Download free assignment PDF https://www.aioumasteracadmey.com

Download free assignment PDF https://www.aioumasteracadmey.com