BHARATHIDASAN UNIVERSITY

TIRUCHIRAPPALLI-620 024,
Tamilnadu, India

Programme : Bachelor of Physical Education

Course Title: Research and Statistics in Physical Education

Course Code: 21BPE43

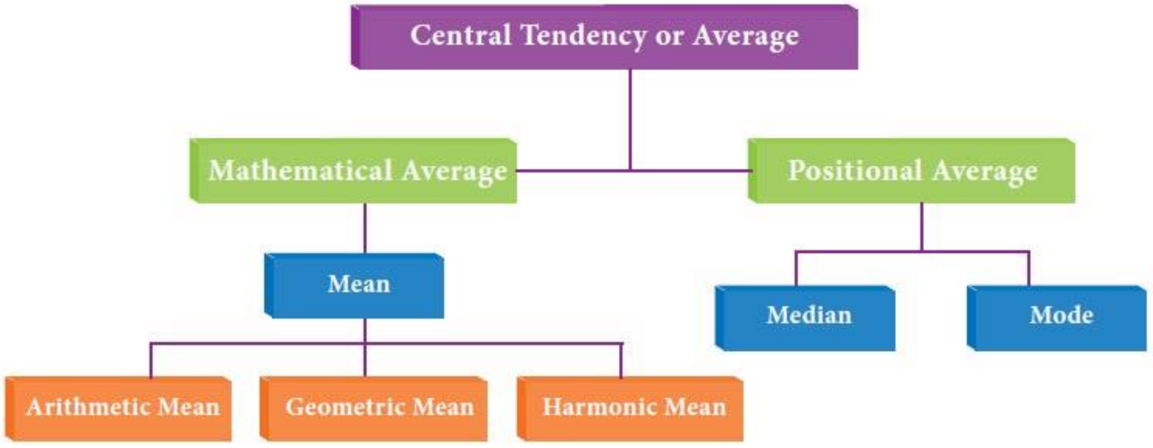
Unit -V

Statistical models in Physical Education and Sports

Dr. R. JAGATHESAN
Guest Lecturer
Department of Physical Education and Yoga

Unit – V

Various measures of central tendency



Mean

In the context of statistics and data analysis, the term "mean" refers to a measure of central tendency. Specifically, it represents the average value of a dataset, providing a single representative value that summarizes the entire set of values.

The **mean**, often called the average, of a numerical set of data, is simply the sum of the data values divided by the number of values. This is also referred to as the arithmetic mean. The mean is the balance point of a distribution.

Mean = sum of the values / the number of values

Step 1: Add the numbers to determine the total number of hours he worked.

24 + 25 + 33 + 50 + 53 + 66 + 78 = 329

Step 2: Divide the total by the number of months.

$$\frac{329}{7} = 47$$

Grouped data

Example

Find the mean of the following data.

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	12	16	6	7	9

Solution:

Class Interval	**Frequency f_i**	**Class Mark x_i**	**($f_i.x_i$)**
0-10	12	5	60

Class Interval	**Frequency f_i**	**Class Mark x_i**	**($f_i.x_i$)**
10-20	16	15	240
20-30	6	25	150
30-40	7	35	245
40-50	9	45	405
	$\sum f_i=50$		$\sum f_{i.}x_i=1100$

$Mean = \sum(f_i.x_i)/\sum f_i = 1100/50 = 22$

The mean, also known as the average, holds significant importance in various aspects of statistics, data analysis, and everyday decision-making. Here are some key reasons why the mean is important:

1. **Central Tendency**: The mean is a measure of central tendency, providing a single representative value that summarizes the entire dataset. It gives an indication of the typical or average value within the data.

2. **Statistical Analysis**: In statistical analysis, the mean is often used to describe the center of a distribution. It helps in understanding the overall pattern and characteristics of the data.

3. **Comparisons**: The mean allows for straightforward comparisons between different datasets or different subsets of the same dataset. For example, comparing the average test scores of students in two different classes or the average monthly sales of a product over two different years.

4. **Decision Making**: Mean values are frequently used in decision-making processes, especially in business, finance, and economics. For instance, determining average revenues, costs, or profits to make strategic decisions.

5. **Data Interpretation**: Mean values provide a quick and easy way to interpret and understand data. They offer a clear point of reference for analyzing trends, patterns, and changes over time.

6. **Sample Representativeness**: In inferential statistics, the mean of a sample is often used as an estimate of the population mean. This is crucial for making inferences and generalizations about the entire population based on a subset of data.

7. **Predictive Modeling**: In predictive modeling and machine learning, the mean can serve as a target variable or feature. Models often aim to predict future mean values or to estimate the mean of an outcome variable based on predictor variables.

8. **Imputation**: The mean can be used to impute missing values in a dataset. While this method has limitations and assumptions, it can help maintain the integrity of the dataset and enable further analysis.

9. **Communicating Results**: Mean values are easily understandable and can effectively communicate key insights to stakeholders, policymakers, and the general public.

Advantages and Disdvantages of Mean

Advantages:

1. **Simple Calculation:** The mean is easy to calculate and understand. It involves adding up all the values in a dataset and then dividing by the total number of values, making it accessible to individuals with varying levels of statistical knowledge.

2. **Reflects Entire Dataset:** The mean takes into account every value in the dataset, providing a comprehensive representation of the data. It considers both the magnitude and direction (positive or negative) of the values.

3. **Statistical Properties:** The mean possesses desirable statistical properties, such as being the point that minimizes the sum of squared deviations from itself (least squares criterion). This property is advantageous in various statistical analyses, including linear regression and analysis of variance (ANOVA).

4. **Useful for Further Analysis:** The mean often serves as a basis for further statistical analysis, such as calculating measures of dispersion (e.g., variance, standard deviation) or conducting hypothesis tests. It provides a starting point for deeper exploration of the dataset's characteristics.

5. **Sampling Distribution:** In statistical inference, the mean of a sample is used to estimate the population mean. The central limit theorem states that the sampling distribution of the mean approaches a normal distribution as the sample size increases, facilitating hypothesis testing and interval estimation.

Disadvantages:

1. **Sensitive to Outliers:** The mean is highly sensitive to extreme values or outliers in the dataset. Even a single outlier can significantly skew the mean, leading to a misrepresentation of the central tendency.

2. **Not Robust for Skewed Distributions:** In datasets with skewed distributions, particularly those with long tails, the mean may not accurately represent the central tendency. It tends to be pulled toward the tail of the distribution, especially when the skewness is substantial.

3. **Not Applicable for Nominal Data:** The mean is not suitable for datasets with nominal or categorical variables where the values lack numerical significance. Attempting to calculate the mean for such data may lead to meaningless results.

4. **Impact of Missing Values:** The presence of missing or incomplete data can affect the accuracy of the mean. Depending on how missing values are handled (e.g., imputation or exclusion), the mean may be biased or unreliable.

5. **Not Always Intuitive:** In some cases, the mean may not be a representative value of the dataset, especially when the distribution is multimodal or contains distinct clusters. In such situations, other measures of central tendency, such as the median or mode, may provide more meaningful insights.

Median

The **median** is the number that falls in the middle position once the data has been organized. Organized data means the numbers are arranged from smallest to largest or from largest to smallest. The median for an odd number of data values is the value that divides the data into two halves. If *n* represents the number of data values

Example

and *n* is an odd number, then the median will be found in the $\frac{n+1}{2}$ position.

Find the median of the following data: 12, 2, 16, 8, 14, 10, 6

Step 1: Organize the data, or arrange the numbers from smallest to largest.

2, 6, 8, 10, 12, 14, 16

Step 2: Since the number of data values is odd, the

Median will be found in the $\frac{n+1}{2}$ position.

$$\frac{n+1}{2} = \frac{7+1}{2} = \frac{8}{2} = 4$$

Step 3: In this case, the median is the value that is found in the fourth position of the organized data.

2, 6, 8, 10, 12, 14, 16

We first arrange the given data values of the observations in ascending order. Then, if n is odd, the median is the (n+1/2). And if n is even, then the median will be the average of the n/2th and the (n/2 +1)th observation. The formula for Calculating Median:

Median, M_e = l + {h x (N/2 – cf)/f}

Where,

- *l = lower limit of median class.*
- *h = width of median class.*
- *f = frequency of median class,*
- *c_f = cumulative frequency of the class preceding the median class.*
- *$N = \sum f_i$*

Calculate the median for the following frequency distribution.

Class Interval	0-8	8-16	16-24	24-32	32-40	40-48
Frequency	8	10	16	24	15	7

Solution:

We may prepare cumulative frequency table as given below,

Class	Frequency	Cumulative Frequency CF
0-8	8	8
8-16	10	18
16-24	16	34
24-32	24	58
32-40	15	73
40-48	7	80
	$N = \sum f_i = 80$	

Now, N = 80 = (N/2) = 40.

The cumulative frequency just greater than 40 is 58 and the corresponding class is 24-32.

Thus, the median class is 24-32.

$l = 24$, $h = 8$, $f = 24$, c_f = c.f. of preceding class = 34, and (N/2) = 40.

Median, M_e = l+ h{(N/2-cf)/f}

$$= 24 + 8 \{(40 - 34)/ 24\}$$

$$= 26$$

*Hence, **median = 26.***

Importance of Median

The median is an important statistical measure for several reasons:

1. **Robustness to Outliers**: Unlike the mean, which can be greatly influenced by extreme values (outliers), the median is resistant to outliers. This makes it a more robust measure of central tendency in datasets with skewed distributions or extreme values.

2. **Representation of Central Value**: The median represents the middle value of a dataset when it's sorted in ascending or descending order. It provides insight into the central value of the dataset, particularly when the distribution is not symmetrical.
3. **Applicability to Ordinal Data**: The median can be calculated for ordinal data, where the values have a natural order but no fixed numerical interpretation. This makes it suitable for analyzing ranked or ordered data, such as rankings or ratings.
4. **Median Income and Wealth Distribution**: In socioeconomic studies, the median income or wealth is often used instead of the mean because it better reflects the typical earnings or assets of a population. This is particularly relevant when income or wealth distributions are highly skewed.
5. **Handling Skewed Distributions**: In datasets with skewed distributions, the median can provide a more accurate representation of the central tendency compared to the mean. This is especially true for positively skewed distributions, where the median is typically less than the mean.
6. **Balance in Group Comparisons**: When comparing groups or populations with different sample sizes or distributions, the median can provide a balanced representation of central tendency. It ensures that each group's central value is given equal weight in the analysis.
7. **Income Inequality and Poverty Analysis**: The median household income is often used as a measure of economic well-being and income distribution within a population. Similarly, the median poverty threshold is used to determine the income level below which a certain percentage of the population falls into poverty.

8. **Survival Analysis**: In medical research and survival analysis, the median survival time is a critical measure for estimating the time until an event of interest (e.g., death, disease recurrence) occurs in a study population.

Advantages and Disadvantages of Median

Advantages:

1. **Robustness to Outliers:** Unlike the mean, which can be heavily influenced by extreme values or outliers, the median is resistant to such effects. It is determined by the middle value of a sorted dataset, making it less sensitive to extreme values and more representative of the central tendency, particularly in skewed distributions.
2. **Suitable for Skewed Distributions:** The median is particularly useful in datasets with skewed distributions, where the mean may not accurately represent the central tendency. It provides a more stable estimate of the typical value, especially when the distribution is heavily skewed or contains outliers.

3. **Simple Interpretation:** The median is easy to interpret and understand. It represents the middle value of a dataset when arranged in ascending or descending order, making it intuitive for individuals with varying levels of statistical knowledge.

4. **Applicable to Ordinal Data:** The median can be calculated for ordinal data, where the values have a natural order but no fixed numerical interpretation. This makes it suitable for analyzing ranked or ordered data, such as preferences or ratings.

5. **Balances Extreme Values:** The median balances out the effects of extreme values when they occur symmetrically around the median. In such cases, the median remains a good representation of the central tendency, providing a more stable estimate than the mean.

Disadvantages:

1. **Less Efficient for Calculations:** Calculating the median involves sorting the dataset, which can be computationally inefficient, especially for large datasets. This process may require additional time and resources compared to calculating the mean.

2. **Less Informational Content:** While the median provides a measure of central tendency, it may not capture as much information about the distribution of the data as the mean does. It focuses solely on the middle value(s) of the dataset without considering the magnitude or direction of the values.

3. **Limited Statistical Properties:** The median lacks certain statistical properties that are desirable in certain analyses, such as being the point that minimizes the sum of squared deviations from itself (as in the least squares criterion). This can limit its applicability in specific statistical procedures or models.

4. **Not Unique in Certain Situations:** In datasets with an even number of observations, the median is not a unique value but rather the average of the two middle values. This lack of uniqueness can complicate interpretations and comparisons, particularly when analyzing small or discrete datasets.

5. **May Mask Distribution Characteristics:** While the median is robust to outliers, it may not fully reflect the distributional characteristics of the data, especially in multimodal distributions or those with distinct clusters. In such cases, other measures of central tendency, such as the mean or mode, may provide additional insights.

Mode

The **mode** of a set of data is simply the value that appears most frequently in the set.

If two or more values appear with the same frequency, each is a mode. The downside to using the mode as a measure of central tendency is that a set of data may have no mode, or it may have more than one mode. However, the same set of data will have only one mean and only one median.

The word modal is often used when referring to the mode of a data set.

- If a data set has only one value that occurs most often, the set is called **unimodal.**
- A data set that has two values that occur with the same greatest frequency is referred to as **bimodal.**
- When a set of data has more than two values that occur with the same greatest frequency, the set is called **multimodal.**

It is that value of a variety that occurs most often. More precisely, the mode is the value of the variable at which the concentration of the data is maximum.

Modal Class: In a frequency distribution, the class having the maximum frequency is called the modal class. The formula for Calculating Mode:

$$M_o = x_k + h\{(f_k - f_{k-1})/(2f_k - f_{k-1} - f_{k+1})\}$$

Where,

- x_k = lower limit of the modal class interval.
- f_k = frequency of the modal class.
- f_{k-1}= frequency of the class preceding the modal class.
- f_{k+1} = frequency of the class succeeding the modal class.
- h = width of the class interval.

Example : Calculate the mode for the following frequency distribution.

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	7	12	28	20	10	10

Solution:

Class 40-50 has the maximum frequency, so it is called the modal class.

$x_k = 40$, $h = 10$, $f_k = 28$, $f_{k-1} = 12$, $f_{k+1} = 20$

Mode, $\mathbf{M_o = x_k + h\{(f_k - f_{k-1})/(2f_k - f_{k-1} - f_{k+1})\}}$

$= 40 + 10\{(28 - 12)/(2 \times 28 - 12 - 20)\}$

$= 46.67$

Hence, **mode = 46.67**

Find the mean, mode, and median for the following data,

Class	0-10	10-20	20-30	30-40	40-50	Total
Frequency	8	16	36	34	6	100

Solution:

Class	**Mid Value x_i**	**Frequency f_i**	**Cumulative Frequency (CF)**	**$f_i . x_i$**
0-10	5	8	8	40
10-20	15	16	24	240
20-30	25	36	60	900

Class	**Mid Value x_i**	**Frequency f_i**	**Cumulative Frequency (CF)**	**$f_i . x_i$**
30-40	35	34	94	1190
40-50	45	6	100	270
		$\sum f_i=100$		$\sum f_i.\ x_i=2640$

Mean **=** $\boldsymbol{\sum(f_i.x_i)/\sum f}$

$= 2640/100$

$= 26.4$

Here, N = 100 ⇒ N / 2 = 50.

Cumulative frequency just greater than 50 is 60 and corresponding class is 20-30.

Thus, the median class is 20-30.

Hence, l = 20, h = 10, f = 36, c = c. f. of preceding class = 24 and N/2=50

Median, M_e = l + h{(N/2 – c_f)/f}

= 20+10{(50-24)/36}

Median = 27.2.

Mode = 3(median) – 2(mean) *= (3 × 27.2 – 2 × 26.4) = 28.8.*

Importance of Mode

The mode, or modal value, is an important statistical measure with several key implications:

1. Identifying Central Tendency: The mode represents the most frequently occurring value or values in a dataset. It provides a clear indication of the central tendency when the data have a clear peak or cluster around specific values.

2. Categorical Data Analysis: In datasets with categorical or nominal variables, such as types of products, colors, or preferences, the mode is often the most relevant measure of central tendency. It indicates the most common category or group within the dataset.

3. Descriptive Statistics: The mode supplements other measures of central tendency, such as the mean and median, by offering additional insights into the distribution of the data. It helps to provide a more comprehensive understanding of the dataset's characteristics.

4. Data Validation: Identifying the mode can be useful for data validation purposes. An unexpected or unusual mode may signal errors in data collection or entry, prompting further investigation.

5. Decision Making: In various fields, including business, marketing, and public policy, the mode is used to inform decision-making processes. For example, in inventory management, knowing the most popular product can guide stocking decisions.

6. Understanding Consumer Behavior: In market research and consumer behavior analysis, identifying the mode helps companies understand consumer preferences and trends. It guides product development, marketing strategies, and resource allocation.

7. Education and Curriculum Development: In educational research, the mode can help identify common learning preferences or teaching methods preferred by students. It informs curriculum development and instructional strategies.

8. Forecasting and Planning: The mode can be valuable for forecasting future trends based on historical data. Recognizing patterns of popularity or preference can aid in predicting future demand or behavior.

9. Multimodal Distributions: Identifying multiple modes in a dataset (multimodal distribution) can reveal complex patterns or subgroups within the data. Understanding these modes can lead to more nuanced insights and targeted interventions.

10. Quality Control: In manufacturing and quality control processes, identifying modes can help identify common defects or issues. It informs efforts to improve product quality and efficiency.

Advantages:

1. Simple Interpretation: The mode is straightforward to interpret as it represents the most frequently occurring value(s) in the dataset. This simplicity makes it accessible and easily understandable, even to individuals with limited statistical knowledge.
2. Applicable to Nominal Data: The mode can be calculated for datasets with nominal or categorical variables, where the values have no inherent numerical significance. It is particularly useful for identifying the most common category or group within the dataset, such as preferred colors or types of products.
3. Robustness to Outliers: Unlike the mean, which can be heavily influenced by extreme values or outliers, the mode is not affected by the magnitude of individual values. It focuses solely on the frequency of occurrence, making it robust to extreme values and skewed distributions.
4. Multimodal Distributions: The mode can identify multiple peaks or modes in a dataset, indicating distinct clusters or subgroups within the data. This ability to detect multimodal distributions provides valuable insights into the underlying structure and patterns of the data.
5. Useful for Discrete Data: In datasets with discrete values, such as counts or frequencies, the mode can provide a meaningful measure of central tendency. It represents the most common value(s) among the discrete observations, facilitating comparisons and decision-making.

Disadvantages:

1. Not Unique: In datasets with uniform distributions or no clear mode, the mode may not be a unique value or may not exist at all. This lack of uniqueness can limit its interpretability and utility, especially in datasets with variability or randomness.
2. Limited Representativeness: The mode may not accurately represent the central tendency of the entire dataset, especially in cases where the distribution is skewed or the dataset contains outliers. It focuses solely on the most frequent value(s), potentially overlooking other important characteristics of the data.
3. Insensitive to Small Changes: The mode may not detect small changes or variations in the dataset, particularly in datasets with large sample sizes or when the frequency of occurrence is evenly distributed across multiple values. It may fail to capture subtle shifts in the underlying distribution.
4. Not Suitable for Interval or Ratio Data: The mode is not applicable to datasets with interval or ratio variables, where the values have meaningful numerical interpretations. Attempting to calculate the mode for such data may lead to misleading results, as it ignores the quantitative differences between values.
5. Difficulty in Handling Continuous Data: While the mode can be calculated for discrete data, it is less straightforward to determine for continuous data, where values are infinitely divisible. In such cases, data discretization or binning may be necessary, potentially leading to loss of information and accuracy.

Calculate the range

The formula to calculate the range is:

$$R = H - L$$

- R = range
- H = highest value
- L = lowest value

The range is the easiest measure of variability to calculate. To find the range, follow these steps:

1. Order all values in your data set from low to high.
2. Subtract the lowest value from the highest value.

This process is the same regardless of whether your values are positive or negative, or whole numbers or fractions.

Range exampleyour data set is the ages of 8 participants.

Participant	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Age	37	19	31	29	21	26	33	36

First, order the values from low to high to identify the lowest value (*L*) and the highest value (*H*).

Age	19	21	26	29	31	33	36	37

Then subtract the lowest from the highest value.

$$\boldsymbol{R} = \boldsymbol{H} - \boldsymbol{L}$$

$$\boldsymbol{R} = \mathbf{37} - \mathbf{19} = \mathbf{18}$$

The range of our data set is **18 years**.

Mean deviation

Is a statistical measure that computes the average deviation from the average value of a given data collection. The mean deviation can be calculated using various data series, such as – continuous data series, discrete data series and individual data series.

$$\textbf{Mean Deviation} = \frac{\Sigma\,|X - \bar{X}|}{N}$$

X = denotes each value in the data set
X¯ = denotes the mean value of the data set
N = total number of data values

Let us look at a simple example to understand the working of the above steps. Suppose we have a dataset **{2, 4, 8, 10}** and we want to calculate the mean deviation about the mean.

Step 1 – We find the mean of the dataset i.e. (2+4+8+10)/4 = 6.

Step 2 – We then subtract each value in the dataset with the mean, get their absolute values i.e.

|2-6| = 4, |4-6| = 2, |8-6| = 2, |10-6| = 4

Step 3 – And add them i.e. 4+2+2+4 = 12.

Step 4 – Finally, we divide this sum by the total number of values in the dataset (4) that will give us the mean deviation. The answer is 12/4 = 3.

Calculate the mean deviation from the median and the co-efficient of mean deviation from the following data:

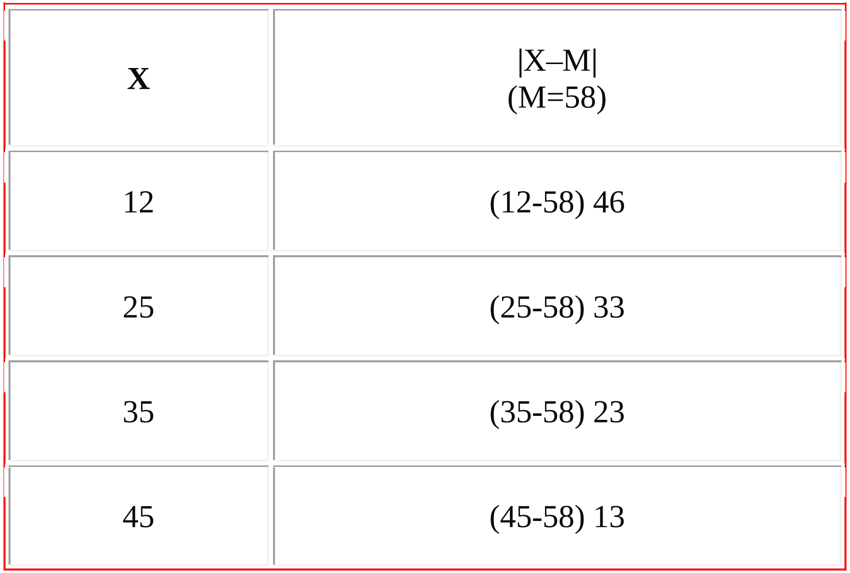
Marks of the students: 86, 25, 87, 65, 58, 45, 12, 71 and 35.

Solution: Arrange the data in ascending order: 12, 25, 35, 45, 58, 65, 71, 86, 87.

Median = Value of the N+1/2

= Value of the 9+1/2=58

Calculation of mean deviation:



X	|X–M| (M=58)
12	(12-58) 46
25	(25-58) 33
35	(35-58) 23
45	(45-58) 13

58	(58-58) 0
65	(65-58) 7
71	(71-58) 13
86	(86-58) 28
87	(87-58) 29
N = 9	∑|X–M|=460

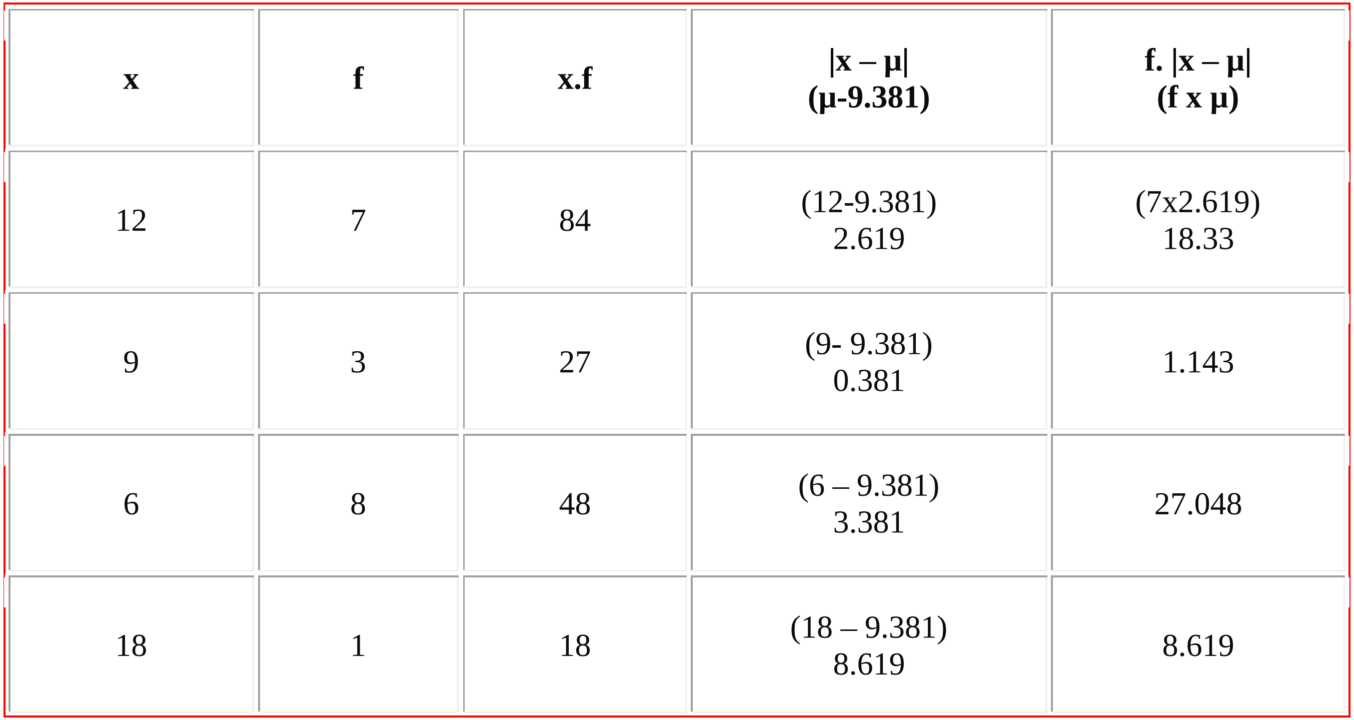
M.D. = ∑|X–M|/N
= 460/9
= 51.11
Co-efficient of Mean Deviation from Median = M.D/M
= 51.11/58

= 0.881

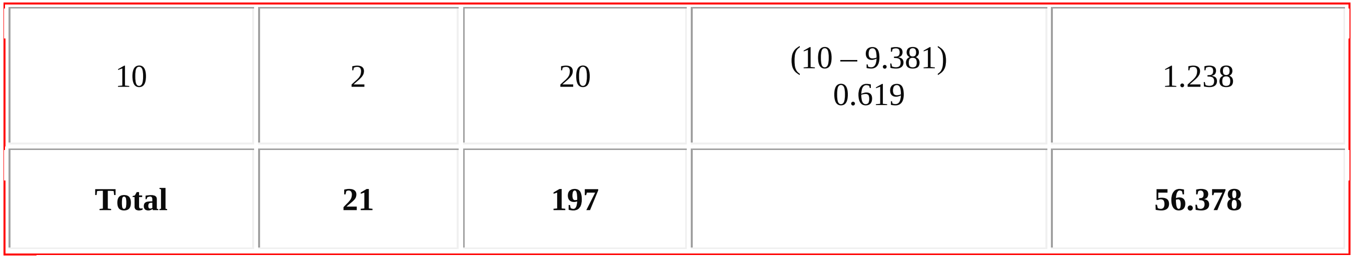
Calculate the mean deviation from mean for the following data.

x	12	9	6	18	10
f	7	3	8	1	2

Answer.



x	**f**	**x.f**	**|x – µ|** **(µ-9.381)**	**f. |x – µ|** **(f x µ)**
12	7	84	(12-9.381) 2.619	(7x2.619) 18.33
9	3	27	(9- 9.381) 0.381	1.143
6	8	48	(6 – 9.381) 3.381	27.048
18	1	18	(18 – 9.381) 8.619	8.619



10	2	20	(10 – 9.381) 0.619	1.238
Total	**21**	**197**		**56.378**

We first find the Mean of the given dataset,

$$\text{Mean } (\mu) = \frac{\sum_1^5 f_i\, x_i}{\sum_1^5 f_i} = \frac{197}{21} = 9.381$$

Finally, we substitute values in the mean deviation about mean formula,

$$\text{Mean Deviation} = \frac{\sum_1^5 f_i |x_i - \mu|}{\sum_1^5 f_i} = \frac{56.378}{21} = 2.684$$

Calculate the mean deviation for the following data.

Class Interval	0 – 2	2 – 4	4 – 6	6 – 8
Frequency	4	2	5	3

Answer.

Class Interval	**Mid-point (x)**	**Frequency (f)**	**f.x**	**|x – μ| = |x – 4|**	**f. |x – μ| (f x μ)**
0 – 2	1	4	4	(1 - 4) 3	(4 x 3) 12
2 – 4	3	2	6	(3 – 4) 1	(2x1)2

4 – 6	5	5	25	(5- 4) 1	(5x1)5
6 – 8	7	3	21	(7 – 4) 3	(3x3)9
Total		**14**	**56**		**28**

Mean (μ) = $\frac{\sum_1^n f_i\, x_i}{\sum_1^n f_i} = \frac{56}{14} = 4$

Finally, we substitute values in the mean deviation formula,

Mean Deviation = $\frac{\sum_1^n f_i |x_i - \mu|}{\sum_1^n f_i} = \frac{28}{14} = 2$

Standard Deviation

Example

The amount of rainfall in a particular season for 6 days are given as 17.8 cm, 19.2 cm, 16.3 cm, 12.5 cm, 12.8 cm and 11.4 cm. Find its standard deviation.

Solution

Arranging the numbers in ascending order we get, 11.4, 12.5, 12.8, 16.3, 17.8, 19.2.

Number of observations $n = 6$

$$\text{Mean} = \frac{11.4+12.5+12.8+16.3+17.8+19.2}{6} = \frac{90}{6} = 15$$

x_i	$d_i = x_i - \overline{x}$ $= x - 15$	d_i^2
11.4	−3.6	12.96
12.5	−2.5	6.25
12.8	−2.2	4.84
16.3	1.3	1.69
17.8	2.8	7.84
19.2	4.2	17.64
		$\Sigma d_i^2 = 51.22$

$$\text{Standard deviation } \sigma = \sqrt{\frac{\Sigma d_i^2}{n}}$$

$$= \sqrt{\frac{51.22}{6}} = \sqrt{8.53}$$

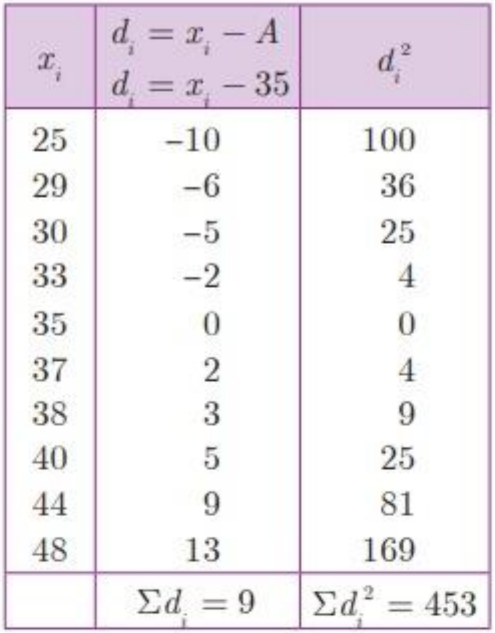
Hence, $\sigma \simeq 2.9$

Example

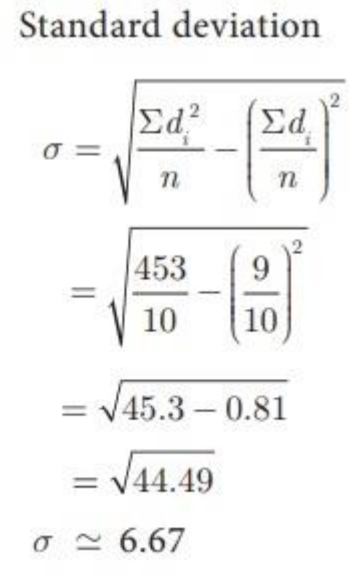
The marks scored by 10 students in a class test are 25, 29, 30, 33, 35, 37, 38, 40, 44, 48. Find the standard deviation.

Solution

The mean of marks is 35.9 which is not an integer. Hence we take assumed mean, $A = 35$, $n = 10$.



x_i	$d_i = x_i - A$ $d_i = x_i - 35$	d_i^2
25	–10	100
29	–6	36
30	–5	25
33	–2	4
35	0	0
37	2	4
38	3	9
40	5	25
44	9	81
48	13	169
	$\Sigma d_i = 9$	$\Sigma d_i^2 = 453$



Standard deviation

$$\sigma = \sqrt{\frac{\Sigma d_i^2}{n} - \left(\frac{\Sigma d_i}{n}\right)^2}$$

$$= \sqrt{\frac{453}{10} - \left(\frac{9}{10}\right)^2}$$

$$= \sqrt{45.3 - 0.81}$$

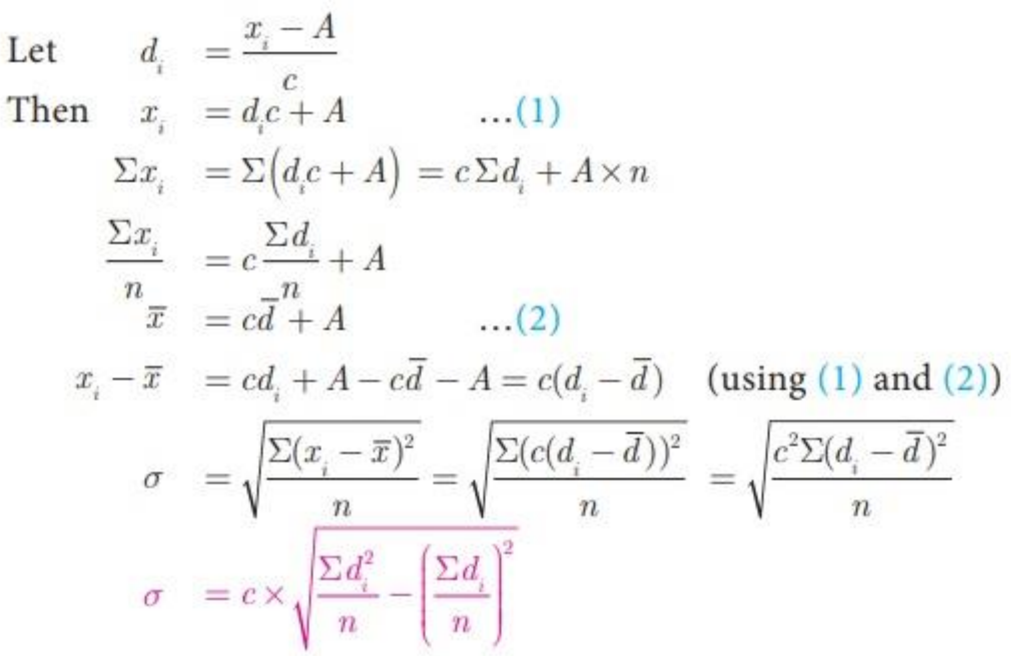
$$= \sqrt{44.49}$$

$$\sigma \simeq 6.67$$

Step deviation method

Let *x* 1 , *x*2, *x* 3 ,...*xn* be the given data. Let A be the assumed mean.

Let *c* be the common divisor of *x i* - *A* .



Let $d_i = \frac{x_i - A}{c}$

Then $x_i = d_i c + A \qquad \ldots(1)$

$$\Sigma x_i = \Sigma\left(d_i c + A\right) = c\Sigma d_i + A \times n$$

$$\frac{\Sigma x_i}{n} = c\frac{\Sigma d_i}{n} + A$$

$$\bar{x} = c\bar{d} + A \qquad \ldots(2)$$

$$x_i - \bar{x} = cd_i + A - c\bar{d} - A = c(d_i - \bar{d}) \quad \text{(using (1) and (2))}$$

$$\sigma = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\Sigma(c(d_i - \bar{d}))^2}{n}} = \sqrt{\frac{c^2\Sigma(d_i - \bar{d})^2}{n}}$$

$$\sigma = c \times \sqrt{\frac{\Sigma d_i^2}{n} - \left(\frac{\Sigma d_i}{n}\right)^2}$$

Example

The amount that the children have spent for purchasing some eatables in one day trip of a school are 5, 10, 15, 20, 25, 30, 35, 40. Using step deviation method, find the standard deviation of the amount they have spent.

Solution We note that all the observations are divisible by 5. Hence we can use the step deviation method. Let the Assumed mean A = 20, n = 8.

x_i	$d_i = x_i - A$ $d_i = x_i - 20$	$d_i = \frac{x_i - A}{c}$ $c = 5$	d_i^2
5	–15	–3	9
10	–10	–2	4
15	–5	–1	1
20	0	0	0
25	5	1	1
30	10	2	4
35	15	3	9
40	20	4	16
		$\Sigma d_i = 4$	$\Sigma d_i^2 = 44$

Standard deviation

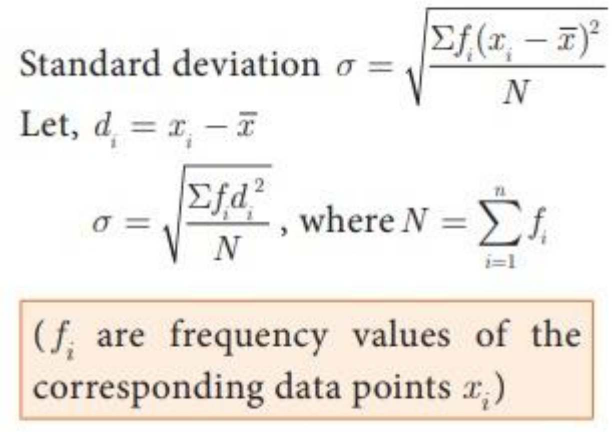
$$\sigma = \sqrt{\frac{\Sigma d_i^2}{n} - \left(\frac{\Sigma d_i}{n}\right)^2} \times c$$

$$= \sqrt{\frac{44}{8} - \left(\frac{4}{8}\right)^2} \times 5 = \sqrt{\frac{11}{2} - \frac{1}{4}} \times 5$$

$$= \sqrt{5.5 - 0.25} \times 5 \ = 2.29 \times 5$$

Calculation of Standard deviation for grouped data

(i) Mean method



Standard deviation $\sigma = \sqrt{\dfrac{\Sigma f_i (x_i - \overline{x})^2}{N}}$

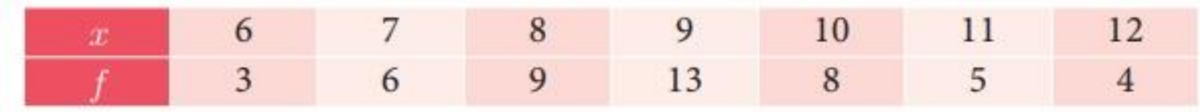
Let, $d_i = x_i - \overline{x}$

$$\sigma = \sqrt{\frac{\Sigma f_i d_i^{\,2}}{N}}, \text{ where } N = \sum_{i=1}^{n} f_i$$

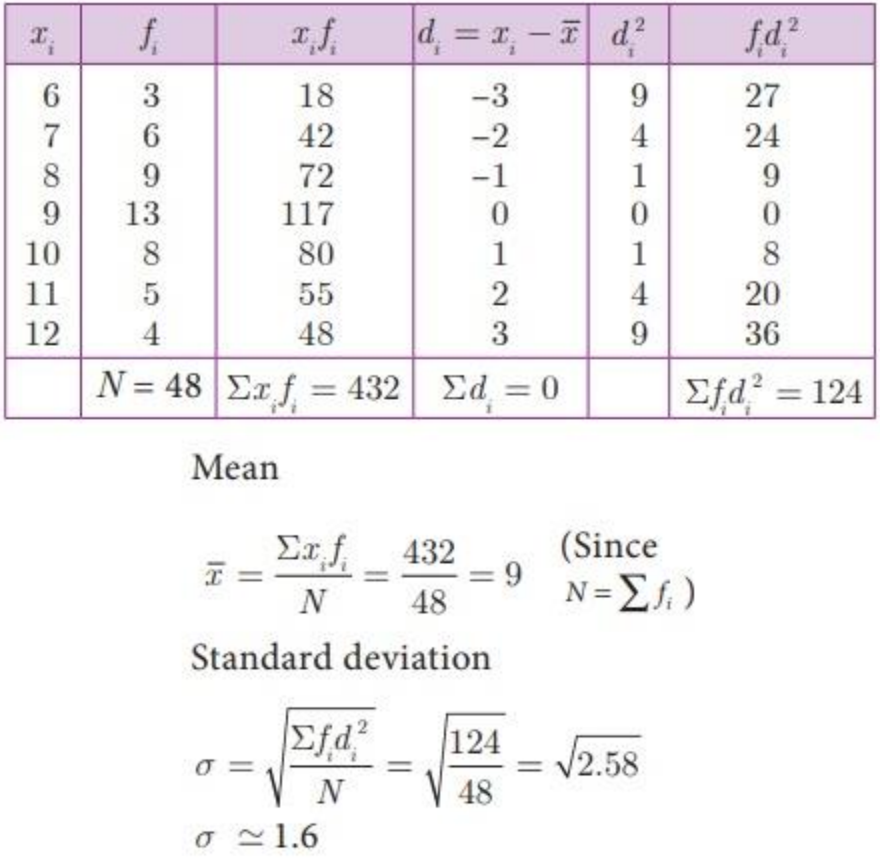
(f_i are frequency values of the corresponding data points x_i)

Example

48 students were asked to write the total number of hours per week they spent on watching television. With this information find the standard deviation of hours spent for watching television.



x	6	7	8	9	10	11	12
f	3	6	9	13	8	5	4



x_i	f_i	$x_i f_i$	$d_i = x_i - \bar{x}$	d_i^2	$f_i d_i^2$
6	3	18	–3	9	27
7	6	42	–2	4	24
8	9	72	–1	1	9
9	13	117	0	0	0
10	8	80	1	1	8
11	5	55	2	4	20
12	4	48	3	9	36
	$N = 48$	$\Sigma x_i f_i = 432$	$\Sigma d_i = 0$		$\Sigma f_i d_i^2 = 124$

Mean

$$\bar{x} = \frac{\Sigma x_i f_i}{N} = \frac{432}{48} = 9 \quad (\text{Since } N = \sum f_i)$$

Standard deviation

$$\sigma = \sqrt{\frac{\Sigma f_i d_i^2}{N}} = \sqrt{\frac{124}{48}} = \sqrt{2.58}$$

$$\sigma \simeq 1.6$$

Solutio

Assumed Mean Method

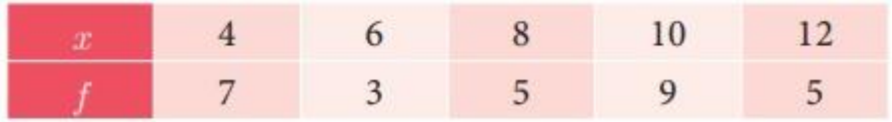
Let *x* 1 , *x*2, *x* 3 , ...*xn* be the given data with frequencies *f*1 , *f*2, *f*3 , ... *fn* respectively. Let *x* be their mean and *A* be the assumed mean..

$$d_i = x_i - A$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\Sigma f_i d_i^2}{N} - \left(\frac{\Sigma f_i d_i}{N}\right)^2}$$

Example 8.12

The marks scored by the students in a slip test are given below.

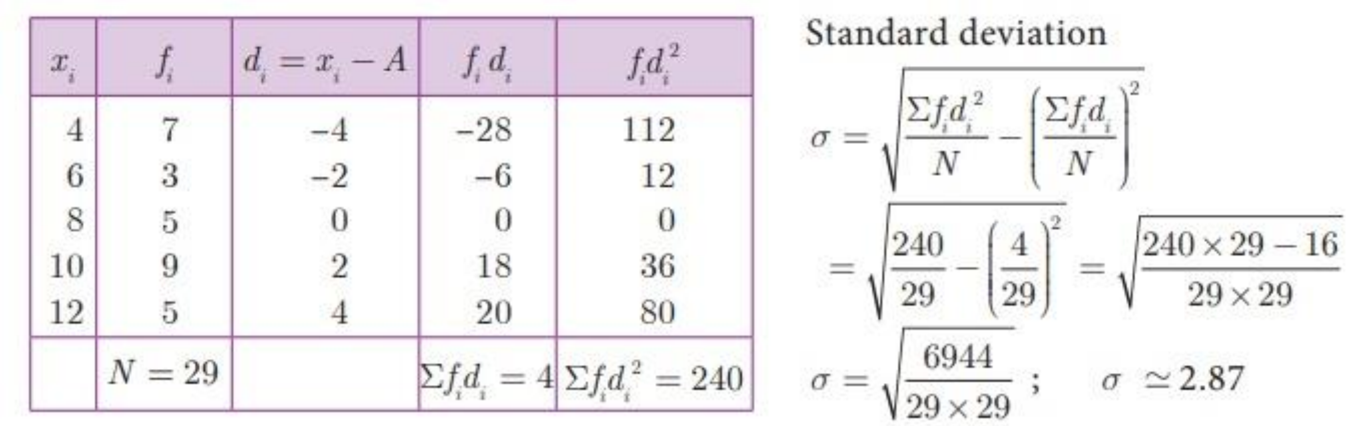


x	4	6	8	10	12
f	7	3	5	9	5

Find the standard deviation of their marks.

Solution

Let the assumed mean, *A* = 8



x_i	f_i	$d_i = x_i - A$	$f_i d_i$	$f_i d_i^2$
4	7	–4	–28	112
6	3	–2	–6	12
8	5	0	0	0
10	9	2	18	36
12	5	4	20	80
	$N = 29$		$\Sigma f_i d_i = 4$	$\Sigma f_i d_i^2 = 240$

Standard deviation

$$\sigma = \sqrt{\frac{\Sigma f_i d_i^2}{N} - \left(\frac{\Sigma f_i d_i}{N}\right)^2}$$

$$= \sqrt{\frac{240}{29} - \left(\frac{4}{29}\right)^2} = \sqrt{\frac{240 \times 29 - 16}{29 \times 29}}$$

$$\sigma = \sqrt{\frac{6944}{29 \times 29}}\ ; \qquad \sigma \simeq 2.87$$