Central Tendency and Dispersion

Md. Aminul Islam Shazid

Outline

1. Introduction
2. Dispersion
3. Outlier
4. Boxplot

Introduction

Central Tendency

- Observations of a variable tend to gather around a single value, this is known as central tendency
- Central tendency is a descriptive measure that represents the center or typical value of a variable
- It provides a summary of the values of the variable

Central Tendency (cont.)

- Mean:
  - Arithmetic mean
  - Geometric mean
  - Harmonic mean
- Median
- Mode

These are different *measures* of central tendency. They represent the "average" value of a dataset in different ways.

Depending on the shape of the distribution and the presence of outliers, different measures are used.

Characteristics of a Good Measure

- Clear and unambiguous definition so that the same data provides the same value of the measure
- Easy to understand and calculate
- Based on all or most of the observations in the sample
- Not unduly affected by outliers so that a few outliers does not distort the result too much
- Representative of the distribution so that the value lies within the range of the data and and describe its central location
- Capable of further mathematical treatment so that it can be used for further analysis

Arithmetic Mean

- The arithmetic mean is the sum of all observations divided by the number of observations
- For a some values $x_1, x_2, \dots, x_n$ of a variable X, the arithmetic mean is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- It uses all observations in the dataset
- The arithmetic mean is easy to compute and interpret
- It is *sensitive* to extreme values (outliers)
- Therefore, it is most appropriate for numerical data that are symmetrically distributed

Example: Arithmetic Mean From Frequency Table

Value, x_i	**Frequency**, f_i	$f_i \cdot x_i$
55	7	385
60	10	600
62	6	372
65	4	260
67	3	201
Total:	30	1818

The mean, $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1818}{30} = 60.6.$

If the data is grouped, then the class midpoints are treated as x_i.

Arithmetic Mean for Grouped Data

Category	**Class midpoint**, x_i	**Frequency**, x_i	$f_i \cdot x_i$
5-30	17.5	7	122.5
30-55	42.5	10	425
55-80	67.5	6	405
80-105	92.5	4	370
105-130	117.5	3	352.5
	Total:	30	1675

Mean = $1675/30 = 55.83$

Weighted Mean

- When caluclating average, sometimes some values may be more important than other values
- In the previous example, the observations appeared different number of times
- Therefore, each value has different level of influence over the center of the distribution
- This is called the weight of each value
- Another example is the calculation of CGPA where the total credit of each semester is the weight of the corresponding GPA

Geometric Mean

- The geometric mean is a measure of central tendency defined as the n-th root of the product of n positive observations
- For positive data $x_1, x_2, \dots, x_n$, the geometric mean is

$$G = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

- It is only defined for positive values
- The geometric mean is appropriate for data involving ratios, rates, or growth factors
- It reduces the influence of very large values compared to the arithmetic mean
- The geometric mean is commonly used for percentage changes and financial returns

Geometric Mean for Grouped Data

- For grouped data, the geometric mean is calculated using class frequencies
- Let $x_1, x_2, \ldots, x_k$ be the class midpoints and $f_1, f_2, \ldots, f_k$ the corresponding frequencies
- The geometric mean is given by

$$G = \left(\prod_{i=1}^{k} x_i^{f_i} \right)^{1/n},$$

  where $n = \sum_{i=1}^{k} f_i$
- In practice, the computation is often simplified using logarithms:

$$\log G = \frac{1}{n} \sum_{i=1}^{k} f_i \log x_i$$

Harmonic mean

- The harmonic mean is a measure of central tendency defined as the reciprocal of the arithmetic mean of reciprocals
- For positive data $x_1, x_2, \ldots, x_n$, the harmonic mean is

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

- It is only defined for positive values
- The harmonic mean gives more weight to smaller observations
- It is appropriate for averaging rates or ratios, such as speeds or densities
- The harmonic mean is strongly affected by very small values

Harmonic Mean for Grouped Data

- For grouped data, the harmonic mean is calculated using class frequencies
- Let $x_1, x_2, \ldots, x_k$ be the class midpoints and $f_1, f_2, \ldots, f_k$ the corresponding frequencies.
- The harmonic mean is given by

$$H = \frac{n}{\sum_{i=1}^{k} \frac{f_i}{x_i}},$$

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

Mode

- The mode is the value that occurs most frequently in a dataset
- A dataset may have:
  - one mode (unimodal),
  - two modes (bimodal), or
  - more than two modes (multimodal)
- The mode can be used for both numerical and categorical data
- A dataset may have no mode if all values occur with the same frequency
- The mode is not affected by extreme values
- For grouped data, the mode is estimated using the modal class

Mode for Grouped Data

$$\text{Mode} = L_0 + \frac{l_1}{l_1 + l_2} \times c,$$

where:

- L_0 is the lower limit of the modal class (class with the highest frequency)
- l_1 is the difference in Frequency between the modal class and the pre-modal class
- l_2 is the difference in Frequency between the modal class and the post-modal class
- c is the class interval

Example: Mode for Grouped Data

Group	Frequency
5-30	7
30-55	10
55-80	6
80-105	4
105-130	3

Mode = $30 + \frac{3}{3+4} \times 30 = 40.71$

Median

- The median is the middle value of a dataset when the observations are arranged in ascending or descending order
- If the number of observations n is odd, the median is the $\frac{n+1}{2}$-th observation
- If n is even, the median is the average of the $\frac{n}{2}$-th and $(\frac{n}{2}+1)$-th observations
- The median divides the dataset into two equal halves
- Therefore, it is the value below which 50% of the data lies
- It is not affected by extreme values (outliers)
- Therefore, it is useful for skewed distributions or data with outliers

Median for Grouped Data

$$\text{Median} = L_m + \frac{\frac{n}{2} - F_c}{f_m} \times c,$$

where:

- L_m = lower limit of the median group, it is the group in which relative cumulative frequency is equal to 0.5 (50%) or the first group in which relative cumulative frequency exceeds 0.5
- n = sample size
- F_c = cumulative frequency of the pre-median class
- f_m = frequency of the median class

Example: Median for Grouped Data

Group	Frequency	Cumulative Frequency
5-30	7	7
30-55	10	17
55-80	6	23
80-105	4	27
105-130	3	30

Here, sample size is 30. Since 50% of 30 is 15, the second group is the median class.

Median = $L_m + \frac{\frac{n}{2} - F_c}{f_m} \times c = 30 + \frac{\frac{30}{2} - 7}{10} \times 25 = 50.$

Trimmed Mean

- The trimmed mean is a measure of central tendency obtained by removing a fixed proportion of the smallest and largest observations
- After trimming, the arithmetic mean is computed using the remaining data
- A $p\%$ trimmed mean removes the lowest $p\%$ and highest $p\%$ of the data
- It is less sensitive to extreme values than the arithmetic mean
- The trimmed mean provides a balance between the mean and the median
- It is useful when the data contain outliers or are moderately skewed

Quantile

- Quantiles are values that divide an ordered dataset into equal parts
- Each part contains the same proportion of observations
- Common quantiles include:
  - Quartiles: divide the data into four equal parts
  - Deciles: divide the data into ten equal parts
  - Percentiles: divide the data into one hundred equal parts
- The median is the second quartile (Q_2) or the 50th percentile, it divides the data in two parts
- Quantiles are useful for describing the distribution and spread of data

Quartile

- Quartiles are values that divide an ordered dataset into four equal parts
- Each part contains approximately 25% of the observations
- The three quartiles are:
    - First quartile (Q_1): 25th percentile
    - Second quartile (Q_2): 50th percentile (the median)
    - Third quartile (Q_3): 75th percentile
- Quartiles are used to describe the spread and position of data
- Sometimes, the minimum value is referred to as the 0th quartile and the maximum value as the 4th quartile

Percentile

- Percentiles divide an ordered dataset into 100 equal parts
- Each percentile represents 1% of the observations
- The p-th percentile is the value below which $p\%$ of the data lie
- The median is the 50th percentile
- Percentiles are widely used in examinations, test scores, and rankings

Someone's IQ score being 90th percentile means that the score is above 90% of the population.

Decile

- Deciles divide an ordered dataset into 10 equal parts
- Each decile represents 10% of the observations
- The k-th decile (D_k) is the value below which $k \times 10\%$ of the data lie
- The fifth decile (D_5) coincides with the median
- Deciles are useful for studying the distribution of data in broader groups

Dispersion

Range

Inter-quartile Range

Used when the range is unduly affected by outliers

Mean Deviation

Standard Deviation

Variance

Coefficient of Variation

Outlier

Boxplot

Questions?