

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 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 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

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

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

Weighted Mean

- When calculating 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, \dots, x_k be the class midpoints and f_1, f_2, \dots, 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, \dots, 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, \dots, x_k be the class midpoints and f_1, f_2, \dots, 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

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

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.

When to Use Mean, Median or Mode

Quantile

Quartile

Percentile

Dispersion

Range

Inter-quartile Range

Mean Deviation

Standard Deviation

Variance

Coefficient of Variation

Outlier

Boxplot

Questions?
