

Chapter 1: Measures of Dispersion

Dr. Anfal A. Alqefari

Contents

1	Introduction	3
2	Population and Sample	3
2.1	Population	3
2.1.1	Properties of a Population	3
2.2	Sample	4
2.2.1	Properties of a Sample	4
2.3	Parameter and Statistic	4
3	Measure of Central Tendency	5
3.1	Arithmetic Mean	5
3.2	Median	6
3.3	Quartiles	6
4	Measures of Dispersion	12
4.1	Range	13
4.2	Interquartile Range	13
4.3	Variance	14
4.4	Standard Deviation	14

4.5	Coefficient of Variation	15
5	Outliers (Extreme Values)	18
5.1	Types of Outliers	19
5.2	Identification of Outliers Using Quartiles	19
5.3	Effect of Outliers on Statistical Measures	20

1 Introduction

Measures of dispersion describe how data values are spread around a central value. In chemical analysis, they are used to evaluate the variation between repeated measurements and to assess the precision and reliability of experimental results. Before discussing measures of dispersion, the most important measures of central tendency, such as the mean, the median, and quartiles, will be introduced. In addition, the difference between a population and a sample will be explained, since different symbols are used to describe each of them in statistical analysis.

2 Population and Sample

In statistics and chemical analysis, data are described using the concepts of population and sample.

2.1 Population

Definition 2.1: Population

A population is the complete set of all possible measurements or observations of interest in a study.

2.1.1 Properties of a Population

- Includes all elements or observations of interest.
- Described by population parameters such as the mean (μ) and variance (σ^2).
- Population parameters are fixed but usually unknown.
- The population size is denoted by N .

2.2 Sample

Definition 2.2: Sample

A sample is a subset selected from the population for measurement and analysis.

2.2.1 Properties of a Sample

- Contains only part of the population.
- Described by sample statistics such as the mean ($\bar{x}$) and variance (s^2).
- Sample statistics are known once the data are collected.
- The sample size is denoted by n .
- An infinite number of different samples can theoretically be drawn from the same population.
- Different samples taken from the same population may give different results.

Remark 2.1

The best sample is a representative sample. A representative sample has characteristics similar to those of the population, such as the mean and the spread of the data. Using a representative sample makes it possible to accurately estimate population parameters.

2.3 Parameter and Statistic

Definition 2.3: Parameter

A parameter is a numerical value that describes a characteristic of a population. Parameters are fixed values but are usually unknown.

Definition 2.4: Statistic

A statistic is a numerical value calculated from sample data. Statistics are used to estimate the corresponding population parameters.

3 Measure of Central Tendency

Measures of central tendency describe the central or typical value of a data set. They summarize data using a single value that represents the center of the distribution. In chemical analysis, measures of central tendency are used to describe the average result of repeated measurements.

The most important measures of central tendency are the mean, the median, and the quartiles.

3.1 Arithmetic Mean

The arithmetic mean is the most commonly used measure of central tendency. It is calculated by dividing the sum of all measured values by the total number of values.

Definition 3.1: Population Mean

The population mean is the average of all values in a population and is given by

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

where x_i represents the i th observation and N is the population size.

Definition 3.2: Sample Mean

The sample mean is the average of all values in a sample and is given by

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

where x_i represents the i th observation and n is the sample size.

3.2 Median

The median is a measure of central tendency that represents the middle value of a data set when the data are arranged in ascending or descending order. It divides the data into two equal parts.

Definition 3.3: Median

The median is the middle value of an ordered data set.

To calculate the median:

- First, arrange the data in ascending order.
- Count the number of observations (n).
- If n is odd, the median is the $\left(\frac{n+1}{2}\right)^{\text{th}}$ observation.
- If n is even, the median is the average of the $\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2} + 1\right)^{\text{th}}$ observations.

Remark 3.1

The median can be calculated for both a population and a sample. It is especially useful when the data contain extreme values, because it is not affected by very large or very small observations.

3.3 Quartiles

Quartiles are measures of position that divide an ordered data set into four equal parts. Each part contains 25% of the data values.

Definition 3.4: Quartiles

Quartiles are values that split an ordered data set into four equal parts.

To calculate quartiles:

- First, arrange the data in ascending order.
- Count the number of observations (n).

- Use the formula

$$Q_i = \left(\frac{i(n+1)}{4} \right)^{\text{th}} \text{ observation, } i = 1, 2, 3$$

Remark 3.2

- **First Quartile (Q_1):** 25% of the data values lie below this value.
- **Second Quartile (Q_2):** 50% of the data values lie below this value (the median).
- **Third Quartile (Q_3):** 75% of the data values lie below this value.

Remark 3.3

Quartiles can also be calculated by dividing the ordered data set into halves.

- Q_1 is the median of the lower half of the data.
- Q_2 is the median of the entire data set.
- Q_3 is the median of the upper half of the data.

This method is commonly used for small data sets and gives results close to those obtained using the formula method.

Remark 3.4

The quartile obtained using the formula method may differ from the quartile obtained using the median of the lower half of the data. The formula method may give a fractional position and requires interpolation, while the lower-half method defines Q_1 as the median of the lower half of the ordered data.

Example 3.1

A student performed five replicate titrations in a chemical analysis experiment. The measured volumes (in mL) are:

10.19, 9.79, 9.69, 10.05, 9.78

Answer the following questions:

1. Calculate the mean.
2. Find the median.
3. Determine the first quartile (Q_1).
4. Determine the second quartile (Q_2).
5. Determine the third quartile (Q_3).

Solution 3.1**1. Mean**

The mean is calculated as

$$\begin{aligned}\bar{x} &= \frac{10.19 + 9.79 + 9.69 + 10.05 + 9.78}{5} \\ &= \frac{49.50}{5} \\ &= 9.9 \text{ mL}\end{aligned}$$

2. Median

First, arrange the data in ascending order:

9.69, 9.78, 9.79, 10.05, 10.19

Second, determine the number of observations:

$$n = 5$$

Since n is odd, the median is the middle value:

$$\text{Median} = 9.79 \text{ mL}$$

3. **First Quartile (Q_1)** First, arrange the data in ascending order:

9.69, 9.78, 9.79, 10.05, 10.19

Here, $n = 5$

$$\begin{aligned} Q_1 &= \left(\frac{i(n+1)}{4} \right)^{\text{th}} \text{ value of the observation} \\ &= \left(\frac{1(5+1)}{4} \right)^{\text{th}} \text{ value of the observation} \\ &= 1.5^{\text{th}} \text{ value of the observation} \\ &= 1^{\text{st}} + 0.5 \left(2^{\text{nd}} - 1^{\text{st}} \right) \\ &= 9.69 + 0.5(9.78 - 9.69) \\ &= 9.735 \text{ mL} \end{aligned}$$

4. **Second Quartile (Q_2)**

First, arrange the data in ascending order:

9.69, 9.78, 9.79, 10.05, 10.19

Using the formula:

$$\begin{aligned}
 Q_2 &= \left(\frac{2(n+1)}{4} \right)^{\text{th}} \text{ value of the observation} \\
 &= \left(\frac{2(5+1)}{4} \right)^{\text{th}} \text{ value of the observation} \\
 &= \left(\frac{12}{4} \right)^{\text{th}} \text{ value of the observation} \\
 &= 3^{\text{rd}} \text{ value of the observation} \\
 &= 9.79 \text{ mL}
 \end{aligned}$$

5. Third Quartile (Q_3)

First, arrange the data in ascending order:

9.69, 9.78, 9.79, 10.05, 10.19

Using the formula:

$$\begin{aligned}
 Q_3 &= \left(\frac{3(n+1)}{4} \right)^{\text{th}} \text{ value of the observation} \\
 &= \left(\frac{3(5+1)}{4} \right)^{\text{th}} \text{ value of the observation} \\
 &= \left(\frac{18}{4} \right)^{\text{th}} \text{ value of the observation} \\
 &= 4.5^{\text{th}} \text{ value of the observation} \\
 &= 4^{\text{th}} + 0.5 (5^{\text{th}} - 4^{\text{th}}) \\
 &= 10.05 + 0.5(10.19 - 10.05) \\
 &= 10.05 + 0.07 \\
 &= 10.12\text{mL}
 \end{aligned}$$

Example 3.2

Consider the results of two chemical experiments.

Data set A (mL): 10.1, 9.8, 10.2, 9.9, 10

Data set B (mL): 10.5, 9, 11, 9.5, 10, 10

Answer the following questions:

1. Calculate the mean for both data sets.
2. Find the median for both data sets.
3. Write an interpretation of the results.

Solution 3.2**1. Mean**

Data set A ($n = 5$):

$$\begin{aligned}\bar{x}_A &= \frac{10.1 + 9.8 + 10.2 + 9.9 + 10}{5} \\ &= \frac{50}{5} \\ &= 10\text{mL}\end{aligned}$$

Data set B ($n = 6$):

$$\begin{aligned}\bar{x}_B &= \frac{10.5 + 9 + 11 + 9.5 + 10 + 10}{6} \\ &= \frac{60}{6} \\ &= 10\text{mL}\end{aligned}$$

2. Median

To find the median, the data must be arranged in ascending order.

Data set A (ordered):

9.8, 9.9, 10, 10.1, 10.2

Since $n = 5$ (odd), the median is the 3rd value:

$$\text{Median}_A = 10$$

Data set B (ordered):

9, 9.5, 10, 10, 10.5, 11

Since $n = 6$ (even), the median is the average of the 3rd and 4th values:

$$\begin{aligned}\text{Median}_B &= \frac{10 + 10}{2} \\ &= 10\end{aligned}$$

3. Interpretation

Both data sets have the same mean and the same median (10.0 mL), even though the number of observations is different.

Remark 3.5

Measures of central tendency alone are not sufficient to fully describe a data set. Different data sets may have the same mean or median but different levels of variation. Measures of dispersion describe how data values spread around the central value and are essential for evaluating precision, consistency, and data reliability in chemical analysis.

4 Measures of Dispersion

In chemical analysis, repeated measurements are often performed to assess the quality and reliability of experimental results. Even when measurements have the same mean value, they may differ in their degree of variability. Measures of dispersion are therefore used to describe how widely repeated measurements are spread around a central value. These measures provide important information

about the precision and consistency of analytical data. In this section, several measures of dispersion are introduced, including the range, the interquartile range, the variance, the standard deviation, and the coefficient of variation.

4.1 Range

The *range* is the simplest measure of dispersion. It is defined as the difference between the maximum and minimum values in a data set. The range provides a quick indication of the spread of repeated measurements; however, it is highly sensitive to extreme values (outliers).

Definition 4.1: Range

For a sample of observations $x_1, x_2, \dots, x_n$, the range is defined as

$$R = \max_{1 \leq i \leq n} (x_i) - \min_{1 \leq i \leq n} (x_i).$$

4.2 Interquartile Range

The *interquartile range* (IQR) is a robust measure of dispersion and is often preferred over the range when data contain extreme values. Unlike the range, which depends only on the minimum and maximum values, the IQR describes the spread of the central 50% of the data and is therefore less sensitive to outliers.

Definition 4.2: Interquartile Range

The interquartile range is defined as the difference between the third quartile and the first quartile:

$$IQR = Q_3 - Q_1,$$

where Q_1 is the first quartile, below which 25% of the data values lie, and Q_3 is the third quartile, below which 75% of the data values lie.

4.3 Variance

The *variance* is a measure of dispersion that describes how far data values are spread around the mean. It is calculated as the average of the squared differences between each observation and the mean. Variance provides a comprehensive measure of data spread, but it is expressed in squared units of the measured quantity.

Definition 4.3: Population Variance

If the true population mean μ is known, the population variance is defined as

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2,$$

where x_i are the population observations and N is the population size.

Definition 4.4: Sample Variance

For a sample of observations $x_1, x_2, \dots, x_n$, the sample variance is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

where $\bar{x}$ is the sample mean and n is the sample size.

4.4 Standard Deviation

The *standard deviation* is a measure of dispersion defined as the square root of the variance. Unlike the variance, the standard deviation is expressed in the same units as the original data, which makes it easier to interpret. It is widely used in chemical analysis to describe the precision and variability of repeated measurements.

Definition 4.5: Population Standard Deviation

If the true population mean μ is known, the population standard deviation is defined as

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2},$$

where x_i are the population observations and N is the population size.

Definition 4.6: Sample Standard Deviation

For a sample of observations $x_1, x_2, \dots, x_n$, the sample standard deviation is defined as

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2},$$

where $\bar{x}$ is the sample mean and n is the sample size.

4.5 Coefficient of Variation

While absolute measures of dispersion such as the variance and standard deviation describe the spread of data, relative measures allow comparison between data sets with different units or scales. The *coefficient of variation* (CV) is a commonly used relative measure of dispersion. It expresses the standard deviation as a percentage of the mean and is especially useful in chemical analysis for comparing precision between different experiments.

Definition 4.7: Coefficient of Variation for a Sample

For a sample of observations, the coefficient of variation is defined as

$$CV = \frac{s}{\bar{x}} \times 100\%,$$

where s is the sample standard deviation and $\bar{x}$ is the sample mean.

Definition 4.8: Coefficient of Variation for a Population

If the true population mean μ is known, the coefficient of variation for the population is defined as

$$CV = \frac{\sigma}{\mu} \times 100\%,$$

where σ is the population standard deviation.

Example 4.1

A student performed five replicate measurements in a chemical analysis experiment. The measured volumes (in mL) are:

10.4, 9.6, 10.1, 9.9, 10

Answer the following:

1. Calculate the mean.
2. Calculate the variance.
3. Calculate the standard deviation.
4. Calculate the coefficient of variation.

Solution 4.1**1. Mean**

$$\begin{aligned}\bar{x} &= \frac{10.4 + 9.6 + 10.1 + 9.9 + 10}{5} \\ &= \frac{50}{5} \\ &= 10 \text{ mL}\end{aligned}$$

2. Variance

$$\begin{aligned}s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\&= \frac{1}{4} [(10.4 - 10)^2 + (9.6 - 10)^2 + (10.1 - 10)^2 \\&\quad + (9.9 - 10)^2 + (10.0 - 10)^2] \\&= \frac{1}{4} (0.16 + 0.16 + 0.01 + 0.01 + 0) \\&= \frac{0.34}{4} \\&= 0.085 \text{ mL}^2\end{aligned}$$

3. Standard Deviation

$$\begin{aligned}s &= \sqrt{s^2} \\&= \sqrt{0.085} \\&= 0.292 \text{ mL}\end{aligned}$$

4. Coefficient of Variation

$$\begin{aligned}CV &= \frac{s}{\bar{x}} \times 100\% \\&= \frac{0.292}{10} \times 100\% \\&= 2.92\%\end{aligned}$$

Remark 4.1

The coefficient of variation (CV) is used to compare the precision of different data sets. A smaller CV means the data are more consistent and more precise, even if the mean values are different.

Example 4.2

A company measures the concentration of a chemical in two batches, Batch A and Batch B. The concentration is measured in mg/L.

The results are:

Batch	Mean (mg/L)	Standard Deviation (mg/L)
A	50	1.5
B	80	2

Compare the coefficient of variation for both batches and decide which batch is more precise.

Solution 4.2

Batch A:

$$\begin{aligned} CV_A &= \frac{1.5}{50} \times 100 \\ &= 3\% \end{aligned}$$

Batch B:

$$\begin{aligned} CV_B &= \frac{2.0}{80} \times 100 \\ &= 2.5\% \end{aligned}$$

Interpretation:

Batch B has a smaller coefficient of variation. This means Batch B is more precise and more consistent than Batch A.

5 Outliers (Extreme Values)

In statistical data analysis, some observations may differ markedly from the rest of the data set. Such observations are referred to as *outliers* or *extreme values*. Outliers represent values that lie far away from the general pattern of the data and may have a significant effect on statistical analysis and interpretation.

Definition 5.1: Outlier

An outlier is an observation that deviates significantly from the other observations in a data set and lies unusually far from the central tendency of the data.

Outliers may occur due to natural variability in the data, experimental or measurement errors, data recording mistakes, or rare but valid events. Therefore, the presence of an outlier does not automatically imply that the value is incorrect; instead, it requires careful investigation.

5.1 Types of Outliers

Outliers can be classified into two main types:

- **High-value outlier (upper extreme):** A value that is much larger than most of the observations in the data set.
- **Low-value outlier (lower extreme):** A value that is much smaller than most of the observations in the data set.

5.2 Identification of Outliers Using Quartiles

One of the most common methods for detecting outliers is based on the interquartile range (IQR).

Definition 5.2: Outlier Detection Using IQR

An observation is considered a potential outlier if it satisfies one of the following conditions:

$$x < Q_1 - 1.5(IQR) \quad \text{or} \quad x > Q_3 + 1.5(IQR),$$

where Q_1 and Q_3 are the first and third quartiles, respectively.

This method is widely used because it is resistant to extreme values and does not depend on the assumption of normality.

5.3 Effect of Outliers on Statistical Measures

Outliers can strongly influence certain statistical measures:

- **Mean:** Highly sensitive (non-resistant) to outliers and may be pulled toward extreme values.
- **Median:** Resistant to outliers and preferred when data are skewed or contain extreme values.

Remark 5.1

In chemical analysis, outliers should not be removed automatically. Each extreme value must be examined to determine whether it results from experimental error, contamination, instrument malfunction, or represents a valid but rare observation.