Self Assignment: Variance and Standard Deviation

2024

Introduction

This document will cover content on variance and standard deviation .

1 Variance

1.1 Definition

Variance measures how far a set of numbers (data points) are spread out from their average value (mean). It quantifies the variability or spread in a dataset.

Given a dataset with n values $x_1, x_2, \ldots, x_n$, the **mean** (average) of the dataset is:

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

The **variance**, denoted by σ^2, is the average of the squared differences from the mean:

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

Here, $(x_i - \mu)^2$ represents the squared difference between each data point and the mean, giving us a measure of how much each point deviates from the average.

1.2 Intuition

Think of variance as a way to measure how much the values in a dataset differ from the average. A low variance means the data points are close to the mean, while a high variance indicates they are spread out over a wider range. For example, the variance of scores in a class with consistent performance will be small, while scores from a more varied class will have a larger variance.

1.3 Sample Variance vs Population Variance

In practice, we often deal with two types of variance:

- **Population Variance:** Used when we have data from the entire population.

- **Sample Variance:** Used when we have a sample from a larger population. For sample variance, the formula is slightly adjusted to account for the fact that we're using a subset of the population:

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

Here, $\bar{x}$ is the sample mean, and dividing by $n-1$ (instead of n) provides an unbiased estimate of the population variance.

2 Standard Deviation

2.1 Definition

The **standard deviation** is the square root of the variance. While variance is measured in squared units, standard deviation brings the units back to the same scale as the data itself, making it easier to interpret.

$$\sigma = \sqrt{\sigma^2}$$

2.2 Intuition

The standard deviation provides a more intuitive understanding of variability. For example, if the standard deviation of a set of exam scores is 5, this tells us that most students scored within 5 points of the average score.

2.3 Properties of Standard Deviation

Standard deviation has several important properties:

- It is always non-negative.
- A standard deviation of zero means all values in the dataset are identical.
- The greater the standard deviation, the more spread out the values are.

3 Application in Computer Science

In computer science, particularly in areas such as **machine learning** and **data analysis**, variance and standard deviation are essential tools:

- **Data Preprocessing:** Standard deviation is used to normalize data, which helps algorithms work more efficiently by ensuring that data features have similar ranges.
- **Loss Functions:** Variance is often minimized in loss functions, which helps machine learning models make better predictions.
- **Anomaly Detection:** By identifying points that are more than a certain number of standard deviations away from the mean, we can detect anomalies or outliers in data.

4 Examples

4.1 Example 1: Calculating Variance and Standard Deviation

Suppose we have the following dataset of exam scores: 75, 80, 85, 90, 95.

Step 1: Calculate the mean:

$$\mu = \frac{75 + 80 + 85 + 90 + 95}{5} = 85$$

Step 2: Calculate the squared differences from the mean:

$$(75-85)^2 = 100, \quad (80-85)^2 = 25, \quad (85-85)^2 = 0, \quad (90-85)^2 = 25, \quad (95-85)^2 = 100$$

Step 3: Calculate the variance:

$$\sigma^2 = \frac{100 + 25 + 0 + 25 + 100}{5} = 50$$

Step 4: Calculate the standard deviation:

$$\sigma = \sqrt{50} \approx 7.07$$

Thus, the variance is 50, and the standard deviation is approximately 7.07.

4.2 Example 2: Sample Variance and Standard Deviation

If the same scores were from a sample, the sample variance would be:

$$s^2 = \frac{100 + 25 + 0 + 25 + 100}{4} = 62.5$$

And the sample standard deviation:

$$s = \sqrt{62.5} \approx 7.91$$

4.3 Example 3 with Scatter Plot

Let us now consider a dataset and visualize it using a scatter plot. We will also indicate the standard deviation on the plot to show how the data is distributed.

Suppose we have the following dataset:

$$(1, 2), (2, 3), (3, 5), (4, 7), (5, 8)$$

Here are the steps to calculate the variance and standard deviation, followed by a scatter plot.

Step 1: Calculate the mean of the y-values:

$$\mu_y = \frac{2 + 3 + 5 + 7 + 8}{5} = 5$$

Step 2: Calculate the variance:

$$\sigma_y^2 = \frac{(2-5)^2 + (3-5)^2 + (5-5)^2 + (7-5)^2 + (8-5)^2}{5} = \frac{9 + 4 + 0 + 4 + 9}{5} = 5.2$$

Step 3: Calculate the standard deviation:

$$\sigma_y = \sqrt{5.2} \approx 2.28$$

Step 4: Scatter Plot:

Below is the scatter plot with the dataset, where we also indicate the standard deviation using a vertical range.

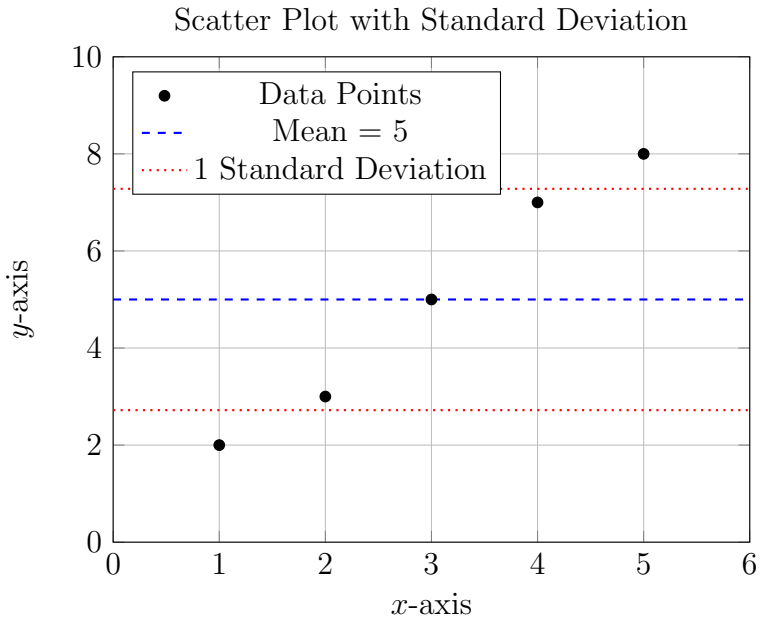

Figure 1: Scatter plot of the dataset with the mean and standard deviation lines.

In the plot, the dashed blue line represents the mean of the y-values (5), and the dotted red lines indicate one standard deviation (approximately 2.28 units) above and below the mean. The data points are generally clustered within this range, illustrating the spread of the dataset.