Unit 3: Measures of the Spread of the Data

3.1 Introduction

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.
- The standard deviation is always positive or zero
- The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread.
- The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

3.2 Real-Life examples

As we know that Standard deviation is a powerful tool used across many fields to understand how spread out data is from its average. Here are a few real-life examples of why standard deviation is important:

3.2.1 Weather Forecasting:

Imagine you're a weather forecaster. City A has an average temperature of 70 degrees Fahrenheit throughout the year, with a low standard deviation. This means temperatures there are predictable and don't fluctuate much. City B, however, has the same average temperature but a high standard deviation. Here, temperatures can vary wildly, making forecasts less reliable.

3.2.2 Quality Control in Manufacturing:

Factories produce thousands of items daily. Standard deviation helps ensure consistency. Let's say a factory makes lightbulbs with an average lifespan of 10,000 hours. A low standard deviation means most bulbs last close to 10,000 hours, indicating good quality control. A high standard deviation suggests a wider range of lifespans, with some bulbs lasting much less or much longer. This can signal production issues.

3.2.3. Investment Decisions:

Investors analyze standard deviation to assess risk. Imagine two stocks, both with an average return of 10% per year. Stock A has a low standard deviation, meaning its returns are usually close to 10%. Stock B has a high standard deviation, indicating returns can fluctuate significantly, sometimes exceeding 10% but also dropping below. Investors seeking stability might choose Stock A, while those comfortable with more risk might consider Stock B.

3.2.4 Educational Assessment:

Teachers can use standard deviation to understand student performance. In a class with a low standard deviation on a test, most students scored close to the average. A high standard deviation suggests scores are spread out, with some students performing much higher or lower than the average. This can help teachers tailor their teaching approach to address the needs of all students.

By understanding standard deviation, you gain valuable insights into the variability of data in many real-life situations. This allows for better decision-making, risk assessment, and overall analysis in various fields.

3.2.5 Plant Growth in a Greenhouse:

Imagine you're a botanist studying tomato plants in a greenhouse. You measure the heights of all the plants at a certain point in their growth cycle. Standard deviation can help you understand the consistency of growth conditions.

- A low standard deviation would indicate that most plants are similar heights, suggesting uniform growing conditions (temperature, light, water) across the greenhouse.
- A high standard deviation would suggest that plant heights vary significantly. This could be due to uneven watering, temperature fluctuations, or other factors affecting different parts of the greenhouse.

3.3 Calculating the Standard Deviation

If x is a number, then the difference "x minus the mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation
To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations**

Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \text{ or } s = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{n-1}}$$

(the first one is used for ungrouped data while the second one is used for grouped data set)
For the sample standard deviation, the denominator is n - 1, where n is sample size

Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} \text{ or } \sigma = \sqrt{\frac{\Sigma f(x-\mu)^2}{N}}$$

(the first one is used for ungrouped data while the second one is used for grouped data set)
For the population standard deviation, the denominator is N, the number of items in the population.

> *"the deviations are measured from the mean and the deviations are squared. In principle, the deviations could be measured from any point, however, our interest is measurement from the center weight of the data"*

3.4 Types of Variability in Samples

When trying to study a population, a sample is often used, either for convenience or because it is not possible to access the entire population. Variability is the term used to describe the differences that may occur in these outcomes. Common types of variability include the following:

- Observational or measurement variability
- Natural variability
- Induced variability
- Sample variability

Here are some examples to describe each type of variability.

3.4.1 Measurement variability

Measurement variability occurs when there are differences in the instruments used to measure or in the people using those instruments. If we are gathering data on how long it takes for a ball to drop from a height by having students measure the time of the drop with a stopwatch, we may experience measurement variability if the two stopwatches used were made by different manufacturers: For example, one stopwatch measures to the nearest second, whereas the other one measures to the nearest tenth of a second. We also may experience measurement variability because two different people are gathering the data. Their reaction times in pressing the button on the stopwatch may differ; thus, the outcomes will vary accordingly. The differences in outcomes may be affected by measurement variability.

3.4.2 Natural variability

Natural variability arises from the differences that naturally occur because members of a population differ from each other. For example, if we have two identical corn plants and we expose both plants to the same amount of water and sunlight, they may still grow at different rates simply because they are two different corn plants. The difference in outcomes may be explained by natural variability.

3.4.3 Induced variability

Induced variability is the counterpart to natural variability; this occurs because we have artificially induced an element of variation (that, by definition, was not present naturally): For example, we assign people to two different groups to study memory, and we induce a variable in one group by limiting the amount of sleep they get. The difference in outcomes may be affected by induced variability.

3.4.4 Sample variability

Sample variability occurs when multiple random samples are taken from the same population. For example, if I conduct four surveys of 50 people randomly selected from a given population, the differences in outcomes may be affected by sample variability.

Example 1

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of n = 20 fifth grade students.

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

$\bar{x}$=$9(1) + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)$ / 20 =**10.525**

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	Freq.	Deviations	Deviations^2	(Freq.)(Deviations^2)
x	f	$(x - \overline{x})$	$(x - \overline{x})^2$	$(f)(x - \overline{x})^2$
9	1	9 – 10.525 = –1.525	$(-1.525)^2 = 2.325625$	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(-1.025)^2 = 1.050625$	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(-0.525)^2 = 0.275625$	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(-0.025)^2 = 0.000625$	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2 = 0.225625$	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2 = 0.950625$	3 × 0.950625 = 2.851875
				The total is 9.7375

Table 3.1

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$$s^2 = \frac{9.7375}{20 - 1} = 0.5125$$

The sample standard deviation s is equal to the square root of the sample variance:

$s = \sqrt{0.5125} = 0.715891$, which is rounded to two decimal places $\boxed{s = 0.72}$

3.5 Explanation of the standard deviation

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. *If you add the deviations, the sum is always zero or close to zero*.

We can begin now by using the standard deviation as a measure of "unusualness."

If somebody asks you that, "How did you do on the test?" you replied "Terrific! Two standard deviations above the mean." This, we will see, is an unusually good exam grade.

The standard deviation, s or σ, is either zero or larger than zero. *Describing the data with reference to the spread is called "variability"*. The variability in data depends upon the method by which the outcomes are obtained; for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are much spread out about the mean; *outliers can make s or σ very large*.

3.6 Standard Deviation for Grouped Frequency Table

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

Mean of Frequency Table $= \dfrac{\sum fm}{\sum f}$; where f = interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how "unusual" individual data is compared to the mean.

Example 2

Find the standard deviation for the data in Table 3.2

Class	Frequency, f	Midpoint, m	$f \cdot m$	$f(m-\overline{x})^2$
0-2	1	1	$1 \cdot 1 = 1$	$1(1-6.88)^2 = 34.57$
3-5	6	4	$6 \cdot 4 = 24$	$6(4-6.88)^2 = 49.77$
6-8	10	7	$10 \cdot 7 = 70$	$10(7-6.88)^2 = 0.14$
9-11	7	10	$7 \cdot 10 = 70$	$7(10-6.88)^2 = 68.14$
12-14	0	13	$0 \cdot 13 = 0$	$0(13-6.88)^2 = 0$
	n=24		$\overline{x} = \dfrac{165}{24} = 6.88$	$s^2 = \dfrac{152.62}{24-1} = 6.64$

Table 3.2

For this data set, we have the mean, x=6.88 and the standard deviation, =2.58. This means that a randomly selected data value would be expected to be 2.58 units from the mean.