3.3-Measures of Variation

Variation:
Variation is a measure of the spread or dispersion of a set of data from its center. Common methods of measuring variation include:
1. Range
2. Standard Deviation
3. Variance

Range:
The range of a set of data values is the difference between the maximum and minimum data values.

\[ \text{Range} = \text{Maximum value} - \text{Minimum Value} \]

Example: Zack has his own business as a painter. The amounts he made in the last five months are shown below.

$ 2416  $ 2423  $ 1644  $ 2036  $ 1267

Find the range.

Solution: The maximum data value is 2423 and the minimum data value is 1267, therefore the range is

\[ \text{Range} = 2423 - 1267 = 1156 \]

Standard Deviation of a Sample:
The standard deviation of a set of sample values, denoted by \( s \), is a measure of variation of values about the mean. The formula for standard deviation is:

\[
s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}
\]

The standard deviation is usually positive. It is only zero if all the data values are the same. It is never negative. The value of the standard deviation \( s \) can increase dramatically with the inclusion of one or more outliers (data values far away from all others).

Example: The instructor of an astronomy course with 4 students often gives quizzes. From the last quiz, the instructor randomly selected the following seven scores: Find the standard deviation:

50  15  31  27  11  42  34

Solution: Because the instructor is using a sample and not all the scores in the course, we use the sample standard deviation formula. To compute the standard deviation, follow the procedure below.
1. Compute the mean.

\[ \bar{x} = \frac{\sum x}{n} = 30 \]

2. Subtract the mean from each sample value and square this value.

\[
\begin{align*}
(50 - 30)^2 &= 400 \\
(15 - 30)^2 &= 225 \\
(31 - 30)^2 &= 1 \\
(27 - 30)^2 &= 9 \\
(11 - 30)^2 &= 361 \\
(42 - 30)^2 &= 144 \\
(34 - 30)^2 &= 16
\end{align*}
\]

3. Add together all the squared values from step 2.

\[ 400 + 225 + 1 + 9 + 361 + 144 + 16 = 1156 \]

4. Divide this sum by n-1 which is one less than the total number of sample values.

\[ \frac{1156}{6} = 192.6 \]

5. Find the square root of the value from step 5.

\[ \sqrt{192.6} = 13.87 \]

When comparing standard deviations of different samples it’s a good practice to compare the sample standard deviations only when the sample means are approximately the same. When the means are different, the standard deviation can give misleading results.

**Standard Deviation of a Population:**

The standard deviation of a population, denoted by \( \sigma \), is a slightly different formula. The formula for standard deviation of a population is:

\[ \sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \]

**Example:** If the 7 quiz scores from the last problem represented all the quiz scores, calculate the standard deviation of the quizzes.

**Solution:** In this problem, \( \bar{x} = \mu \) and \( N = 7 \).

\[ \sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} = \sqrt{\frac{1156}{7}} = 12.85 \]
Variance of a Sample and a Population:
The variance of a set of values is a measure of variation equal to the square of the standard deviation. The sample variance $s^2$ is an unbiased estimator of the population variance $\sigma^2$ which means that the values of $s^2$ tend to target the value of $\sigma^2$ instead of systematically overestimating or underestimating $\sigma^2$.

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \quad \text{Variance of a sample}$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} \quad \text{Variance of a Population}$$

Standard Deviation vs. Variance
Both standard deviation and variance measure dispersion from center of a set of data and each has special value in different applications. Both will be used throughout this course.

Caution on Units:
The value of standard deviation will have the same units as the original measure. Variation, however, will be in square units.

For instance, if we are finding the standard deviation of units of time in minutes, the standard deviation will be in minutes. The variance however will be expressed in square minutes which are an abstract notion.

Range Rule of Thumb:
A simple tool for understanding standard deviation is the range rule of thumb. This rule is based on the principle that for many data sets 95% of the data values lie within 2 standard deviations of the mean. This rule may be expressed as:

Minimum Value = Mean - 2 times Standard Deviation
Maximum Value = Mean - 2 times Standard Deviation

The standard deviation may be approximated using this rule as follows:

$$s \approx \frac{\text{Range}}{4}$$

Example: The heights in feet of people who work in an office are as follows.

5.8  6.1  5.9  5.4  5.6  5.8  5.9  6.2  6.1  5.8

Use the range rule of thumb to estimate the standard deviation.

Solution: The maximum value is 6.2 and the minimum value is 5.4, therefore we have

$$s = \frac{6.2 - 5.4}{4} = 0.2$$
Empirical Rule:
Another concept that is helpful in interpreting the value of standard deviation is the empirical rule. This rule states that for data sets having a normal distribution (bell-shaped), the following properties apply:

About 68% of the data values fall within 1 standard deviation of the mean.

About 95% of the data values fall within 2 standard deviations of the mean.

About 99.7% of the data values fall within 3 standard deviations of the mean.
Example: The systolic blood pressure of 18-year-old women is normally distributed with a mean of 120 mmHg and a standard deviation of 12 mmHg. What percentage of 18-year-old women have a systolic blood pressure between 96 mmHg and 144 mmHg?

Solution: Find the range of values within 1, 2 and 3 standard deviations of the mean.

1 SD: 120 ± 12 68% of the values are between 108 and 132

2 SD: 120 ± 2(12) 95% of the values are between 96 and 144

3 SD: 120 ± 3(12) 99.7% of the values are between 84 and 156

Chebyshev’s Theorem:
A third concept that is helpful in understanding or interpreting a value of a standard deviation is Chebyshev’s Theorem. The empirical rule applies only to data sets with bell-shaped distributions, but Chebyshev’s theorem applies to any data set. Unfortunately, the results from Chebyshev’s theorem are only approximate and the theorem is therefore limited in its usefulness.

Chebyshev’s Theorem:
The proportion (or fraction) of any set of data lying within $K$ standard deviations of the mean is always at least $1 - \frac{1}{K^2}$, where $K$ is any positive number greater than 1.

- For $K = 2$, at least 3/4 (or 75%) of all values lie within 2 standard deviations of the mean.
- For $K = 3$, at least 8/9 (or 89%) of all values lie within 3 standard deviations of the mean.

Example: The heights of the adults in one town have a mean of 67.5 inches and a standard deviation of 3.4 inches. What can you conclude from Chebyshev's theorem about the percentage of adults in the town whose heights are between 57.3 and 77.7 inches?

Solution: Because $67.5 + 3(3.4) = 77.7$ and $67.5 - 3(3.4) = 57.3$ we can conclude that approximately 89% of the adults in the town will have heights between these values.

Properties of the Standard Deviation:

- Measures the variation among data values
- Values close together have a small standard deviation, but values with much more variation have a larger standard deviation
- Has the same units of measurement as the original data
- For many data sets, a value is unusual if it differs from the mean by more than two standard deviations
- Compare standard deviations of two different data sets only if they use the same scale and units, and they have means that are approximately the same
**Coefficient of Variation:**
The coefficient of variation is used for comparing variation in different sets of data where either the means are significantly different or the units are of a different scale or measurement.

Coefficient of variation for a sample: \[ CV = \frac{s}{\bar{x}} \cdot 100\% \]

Coefficient of variation for a population: \[ CV = \frac{\sigma}{\mu} \cdot 100\% \]

If one population or sample has a higher coefficient of variation than the other, then we can conclude that the sample or population has more variation.

**Example:** The customer service department of a phone company is experimenting with two different systems. On Monday they try the first system which is based on an automated menu system. On Tuesday they try the second system in which each caller is immediately connected with a live agent. A quality control manager selects a sample of seven calls each day. He records the time for each customer to have his or her question answered. The times (in minutes) are listed below.

- Automated Menu: 11.7, 7.5, 3.9, 2.9, 9.2, 6.3, 5.5
- Live agent: 6.4, 2.8, 4.4, 4.1, 3.4, 5.2, 3.7

Use the coefficient of variation to compare the variation in the amount of time it takes for a customer to have their questions answered between the two systems.

**Solution:**
Automated Menu: 47.1%
Live agent: 29.1%
There is substantially more variation in the times for the automated menu system.