Chapter 3: Data Description

Diana Pell

Section 3.2: Measures of Variation

Exercise 1. A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

<table>
<thead>
<tr>
<th>Brand A</th>
<th>Brand B</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>35</td>
</tr>
<tr>
<td>60</td>
<td>45</td>
</tr>
<tr>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
</tr>
</tbody>
</table>

Note: The means are the same for both brands, the spread, or variation, is quite different. For the spread or variability of a data set, three measures are commonly used: range, variance, and standard deviation.
Range

\[ R = \text{highest value} - \text{lowest value} \]

**Exercise 2.** Find the ranges for the paints in Exercise 1.

**Note:** Outliers can strongly affect the range.

**Population Variance and Standard Deviation**

**Exercise 3.** Find the variance and standard deviation for the data set for brand A paint in Exercise 1.
The **Population variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is \( \sigma^2 \) (\( \sigma \) is the Greek lowercase letter sigma).

The formula for the population variance is

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

where \( X \) = individual value, \( \mu \) = population mean, and \( N \) = population size.

The **population standard deviation** is the square root of the variance. The symbol for the population standard deviation is \( \sigma \).

The corresponding formula for the population standard deviation is

\[
\sigma = \sqrt{\sigma^2}
\]

**Note:** The variance is actually the average of the square of the distance that each value is from the mean. Therefore, if the values are near the mean, the variance will be small. In contrast, if the values are far from the mean, the variance will be large.

**Exercise 4.** Find the variance and standard deviation for brand B paint data in Exercise 1.

What brand has more variability?
Sample Variance and Standard Deviation

The formula for the **sample variance** (denoted by $s^2$) is

$$ s^2 = \frac{\sum(X - \bar{X})^2}{n - 1} $$

The corresponding formula for the sample standard deviation is

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

**Exercise 5.** The number of public school teacher strikes in Pennsylvania for a random sample of school years is shown. Find the sample variance and the sample standard deviation.

$$ 9 \ 10 \ 14 \ 7 \ 8 \ 3 $$
Shortcut formulas for computing $s^2$ and $s$

Variance

$$s^2 = \frac{n \cdot (\sum X^2) - (\sum X)^2}{n(n - 1)}$$

and standard deviation: $s = \sqrt{s^2}$

**Exercise 6.** The number of public school teacher strikes in Pennsylvania for a random sample of school years is shown. Find the sample variance and sample standard deviation.

9 10 14 7 8 3
Variance for Grouped Data

\[ s^2 = \frac{n \cdot (\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n - 1)} \]

**Exercise 7.** Find the sample variance and the sample standard deviation for the frequency distribution below. The data represent the number of miles that 20 runners ran during one week.

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Midpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5–10.5</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>10.5–15.5</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>15.5–20.5</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>20.5–25.5</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>25.5–30.5</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>30.5–35.5</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>35.5–40.5</td>
<td>2</td>
<td>38</td>
</tr>
</tbody>
</table>
The **coefficient of variation** denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples: \[ \text{CVar} = \frac{s}{\bar{X}} \times 100 \]

For populations: \[ \text{CVar} = \frac{\sigma}{\mu} \times 100 \]

**Exercise 8.** The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

---

**Range Rule of Thumb**

A rough estimate of the standard deviation is

\[ s \approx \frac{\text{range}}{4} \]

**Note:** The range rule of thumb is only an approximation and should be used when the distribution of data values is unimodal and roughly symmetric.

The range rule of thumb can be used to estimate the largest and smallest data values of a data set.

**Chebyshev’s Theorem**

The proportion of values from a data set that will fall within \( k \) standard deviations of the mean will be at least \( 1 - \frac{1}{k^2} \), where \( k \) is a number greater than 1 (\( k \) is not necessarily an integer).

* At least three-fourths, or 75%, of all data values fall within 2 standard deviations of the mean.
* At least eight-ninths, or 89%, of all data values fall within 3 standard deviations of the mean.
Exercise 9. The mean price of houses in a certain neighborhood is $50,000, and the standard deviation is $10,000. Find the price range for which at least 75% of the houses will sell.

Exercise 10. A survey of local companies found that the mean amount of travel allowance for couriers was $0.25 per mile. The standard deviation was $0.02. Using Chebyshev’s theorem, find the minimum percentage of the data values that will fall between $0.20 and $0.30.

The Empirical (Normal) Rule

When a distribution is bell-shaped (or what is called normal), the following statements, which make up the empirical rule, are true.

* Approximately 68% of the data values will fall within 1 standard deviation of the mean.
* Approximately 95% of the data values will fall within 2 standard deviations of the mean.
* Approximately 99.7% of the data values will fall within 3 standard deviations of the mean.
**Exercise 11.** The table lists means and standard deviations. The mean is the number before the plus/minus, and the standard deviation is the number after the plus/minus. The results are from a study attempting to find the average blood pressure of older adults. Use the results to answer the questions.

<table>
<thead>
<tr>
<th></th>
<th>Normotensive</th>
<th>Hypertensive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men (n = 1200)</td>
<td>Women (n = 1400)</td>
</tr>
<tr>
<td>Age</td>
<td>55 ± 10</td>
<td>55 ± 10</td>
</tr>
<tr>
<td>Blood pressure (mm Hg)</td>
<td>123 ± 9</td>
<td>121 ± 11</td>
</tr>
<tr>
<td>Systolic</td>
<td>78 ± 7</td>
<td>76 ± 7</td>
</tr>
<tr>
<td>Diastolic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Apply Chebyshev’s theorem to the systolic blood pressure of normotensive men. At least how many of the men in the study fall within 1 standard deviation of the mean?

2. At least how many of those men in the study fall within 2 standard deviations of the mean?

3. Assume that blood pressure is normally distributed among older adults. Answer the following questions, using the empirical rule instead of Chebyshev’s theorem. What are the ranges for the diastolic blood pressure (normotensive and hypertensive) of older women? Do the normotensive, male, systolic blood pressure ranges overlap with the hypertensive, male, systolic blood pressure ranges?