Chapter 3
Statistics for Describing, Exploring, and Comparing Data

3-1 Overview
3-2 Measures of Center
3-3 Measures of Variation

Section 3-1: Overview

- **Descriptive Statistics**
  summarize or describe the important characteristics of a known set of data

- **Inferential Statistics**
  use sample data to make inferences (or generalizations) about a population

---

3-2 Measures of Center

**Key Concept**
When describing, exploring, and comparing data sets, these characteristics are usually extremely important: center, variation, distribution, outliers, and changes over time.

---

**Definition**

- **Measure of Center**
  the value at the center or middle of a data set

---

**Definition**

- **Arithmetic Mean**
  (Mean)

  the measure of center obtained by adding the values and dividing the total by the number of values

---

**Notation**

- $\Sigma$ denotes the sum of a set of values.
- $x$ is the variable usually used to represent the individual data values.
- $n$ represents the number of values in a sample.
- $N$ represents the number of values in a population.
**Notation**

\[ \bar{x} \text{ is pronounced ‘x-bar’ and denotes the mean of a set of sample values} \]
\[ \bar{x} = \frac{\sum x}{n} \]

\[ \mu \text{ is pronounced ‘mu’ and denotes the mean of all values in a population} \]
\[ \mu = \frac{\sum x}{N} \]

**Definitions**

- **Median**
  - the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude
  - often denoted by \( \bar{x} \) (pronounced ‘x-tilde’)
  - is not affected by an extreme value

**Finding the Median**

- If the number of values is odd, the median is the number located in the exact middle of the list.

- If the number of values is even, the median is found by computing the mean of the two middle numbers.

**Mode - Examples**

- **Mode**
  - the value that occurs most frequently
  - Mode is not always unique
  - A data set may be:
    - Bimodal
    - Multimodal
    - No Mode

Mode is the only measure of central tendency that can be used with nominal data.
Definition

- **Midrange**
  
  the value midway between the maximum and minimum values in the original data set

\[
\text{Midrange} = \frac{\text{maximum value} + \text{minimum value}}{2}
\]

Round-off Rule for Measures of Center

Carry one more decimal place than is present in the original set of values.

Mean from a Frequency Distribution

Assume that in each class, all sample values are equal to the class midpoint.

\[
\bar{x} = \frac{\sum (f \cdot x)}{\sum f}
\]

Weighted Mean

In some cases, values vary in their degree of importance, so they are weighted accordingly.

\[
\bar{x} = \frac{\sum (w \cdot x)}{\sum w}
\]

Best Measure of Center

- **Mean**
  - Find the sum of all values, then divide by the number of values
- **Median**
  - The middle value in the ordered data
- **Mode**
  - The value that occurs most frequently
- **Midrange**
  - The value midway between the maximum and minimum values
Definitions

- **Symmetric**
  - Distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half.

- **Skewed**
  - Distribution of data is skewed if it is not symmetric and if it extends more to one side than the other.

Skewness

- In this section we have discussed:
  - Types of measures of center
    - Mean
    - Median
    - Mode
  - Mean from a frequency distribution
  - Weighted means
  - Best measures of center
  - Skewness

Section 3-3: Measures of Variation

**Key Concept**

Because this section introduces the concept of variation, which is something so important in statistics, this is one of the most important sections in the entire book.

Place a high priority on how to interpret values of standard deviation.

Definition

The **range** of a set of data is the difference between the maximum value and the minimum value.

\[ \text{Range} = (\text{maximum value}) - (\text{minimum value}) \]

Definition

The **standard deviation** of a set of sample values is a measure of variation of values about the mean.
Sample Standard Deviation Formula

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

Sample Standard Deviation (Shortcut Formula)

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

Standard Deviation - Important Properties

- The standard deviation is a measure of variation of all values from the mean.
- The value of the standard deviation \( s \) is usually positive.
- 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).
- The units of the standard deviation \( s \) are the same as the units of the original data values.

Population Standard Deviation

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

This formula is similar to the previous formula, but instead, the population mean and population size are used.

Definition

- The variance of a set of values is a measure of variation equal to the square of the standard deviation.

Sample variance: Square of the sample standard deviation \( s \)

Population variance: Square of the population standard deviation \( \sigma \)

Variance - Notation

standard deviation squared

Notation \( \begin{cases} s^2 & \text{Sample variance} \\ \sigma^2 & \text{Population variance} \end{cases} \)
**Round-off Rule for Measures of Variation**

Carry one more decimal place than is present in the original set of data.

Round only the final answer, not values in the middle of a calculation.

**Estimation of Standard Deviation**

**Range Rule of Thumb**

For estimating a value of the standard deviation $s$, use

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

Where range = (maximum value) – (minimum value)

**Empirical (68-95-99.7) Rule**

For data sets having a distribution that is approximately bell shaped, the following properties apply:

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviations of the mean.
- About 99.7% of all values fall within 3 standard deviations of the mean.

**The Empirical Rule**

- 68% within 1 standard deviation
- 95% within 2 standard deviations
- 99.7% within 3 standard deviations
The Empirical Rule

99.7% of data are within 3 standard deviations of the mean ($\mu - 3\sigma$ to $\mu + 3\sigma$).

95% within 2 standard deviations

68% within 1 standard deviation

The end of Section 3-3 has a detailed explanation of why $n - 1$ rather than $n$ is used. The student should study it carefully.

Definition

Chebyshev’s Theorem

The proportion (or fraction) of any set of data lying within $K$ standard deviations of the mean is always at least $1 - 1/K^2$, 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.

Rationale for using $n-1$ versus $n$

Recap

In this section we have looked at:

- Range
- Standard deviation of a sample and population
- Variance of a sample and population
- Range rule of thumb
- Empirical distribution
- Chebyshev’s theorem