Chapter 3
Describing, Exploring, and Comparing Data

A) Measures of Center
B) Measures of Variation
C) Measures of Relative Standing
D) Exploratory Data Analysis

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

Σ denotes the addition 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

⃗x is pronounced ‘x-bar’ and denotes the mean of a set of sample values

⃗x = Σx / n

μ is pronounced ‘mu’ and denotes the mean of all values in a population

μ = Σx / N

Definitions

Median
the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude

often denoted by ⃗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.

Examples

<table>
<thead>
<tr>
<th>5.40</th>
<th>1.10</th>
<th>0.42</th>
<th>0.73</th>
<th>0.48</th>
<th>1.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.42</td>
<td>0.48</td>
<td>0.73</td>
<td>1.10</td>
<td>1.10</td>
<td>5.40</td>
</tr>
</tbody>
</table>

**even number of values – no exact middle shared by two numbers**

\[
\frac{0.73 + 1.10}{2} = \text{MEDIAN is 0.915}
\]

<table>
<thead>
<tr>
<th>5.40</th>
<th>1.10</th>
<th>0.42</th>
<th>0.73</th>
<th>0.48</th>
<th>1.10</th>
<th>0.66</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.42</td>
<td>0.48</td>
<td>0.66</td>
<td>0.73</td>
<td>1.10</td>
<td>1.10</td>
<td>5.40</td>
</tr>
</tbody>
</table>

**odd number of values**

MEDIAN is 0.73

Definitions

- **Mode**: the value that occurs most frequently.
  - The mode is not always unique. A data set may be:
    - Bimodal
    - Multimodal
    - No Mode
  - denoted by M
  - the only measure of central tendency that can be used with nominal data

Examples

- a. 5.40 1.10 0.42 0.73 0.48 1.10
  - **Mode is 1.10**
- b. 27 27 27 55 55 55 88 88 99
  - **Bimodal - 27 & 55**
- c. 1 2 3 6 7 8 9 10
  - **No Mode**

Definitions

- **Midrange**: the value midway between the highest and lowest values in the original data set.

\[
\text{Midrange} = \frac{\text{highest score} + \text{lowest score}}{2}
\]

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

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

**Definitions**

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

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

**Best Measure of Center**

<table>
<thead>
<tr>
<th>Measure of Center</th>
<th>Definition</th>
<th>Easy to Compute?</th>
<th>Good Connecting?</th>
<th>Sensitive?</th>
<th>Takes Entry Values into Account?</th>
<th>Effected by Outliers?</th>
<th>Advantages and Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (\bar{x} = \frac{\sum x}{n})</td>
<td>mean, center, “average”</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>either</td>
<td>used throughout this book, works well with many statistical methods. Not a good choice if there are some outliers. Most appropriate for data at the normal level.</td>
</tr>
<tr>
<td>Median</td>
<td>middle value</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>most frequent data value</td>
<td>sometimes used</td>
<td>yes</td>
<td>sometimes used</td>
<td>yes</td>
<td>low</td>
<td></td>
</tr>
</tbody>
</table>

**Skewness**

- **Symmetric**

- **Skewed to the Left (Negatively)**

- **Skewed to the Right (Positively)**
Measures of Variation

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

Definition

The range of a set of data is the difference between the highest value and the lowest value:

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

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)

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

Standard Deviation - Key Points

- 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 Formula 2-4, 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

- \( S^2 \): Sample variance
- \( \sigma^2 \): Population variance

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.

Definition

The coefficient of variation (or CV) for a set of sample or population data, expressed as a percent, describes the standard deviation relative to the mean.

Sample

\[ CV = \frac{S}{\bar{x}} \times 100\% \]

Population

\[ CV = \frac{\sigma}{\mu} \times 100\% \]

Standard Deviation from a Frequency Distribution

Formula 2-6

\[ S = \sqrt{\frac{n \left[ \sum (f \cdot x^2) \right] - \left[ \sum (f \cdot x) \right]^2}{n (n - 1)}} \]

Use the class midpoints as the x values.
Estimation of Standard Deviation
Range Rule of Thumb

For estimating a value of the standard deviation \( s \), use

\[
\frac{\text{Range}}{4}
\]

Where range = (highest value) – (lowest value)

---

Estimation of Standard Deviation
Range Rule of Thumb

For interpreting a known value of the standard deviation \( s \), find rough estimates of the minimum and maximum “usual” values by using:

Minimum “usual” value \( \approx (\text{mean}) - 2 \times (\text{standard deviation}) \)

Maximum “usual” value \( \approx (\text{mean}) + 2 \times (\text{standard deviation}) \)

---

Definition

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

FIGURE 2-13

---

The Empirical Rule

FIGURE 2-13
**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 \( 1 - \frac{1}{K^2} \), 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

**Rationale for Formula 2-4**
The end of Section 2-5 has a detailed explanation of why Formula 2-4 is employed instead of other possibilities and, specifically, why \( n - 1 \) rather than \( n \) is used. The student should study it carefully.

**Definition**

- **z Score** (or standard score)
  the number of standard deviations that a given value \( x \) is above or below the mean.

**Measures of Position**

- **Sample**
  \[ z = \frac{x - \bar{x}}{s} \]
  Round to 2 decimal places

- **Population**
  \[ z = \frac{x - \mu}{\sigma} \]

**Interpreting Z Scores**

- **Q₁ (First Quartile)** separates the bottom 25% of sorted values from the top 75%.
- **Q₂ (Second Quartile)** same as the median; separates the bottom 50% of sorted values from the top 50%.
- **Q₃ (Third Quartile)** separates the bottom 75% of sorted values from the top 25%.
**Quartiles**

\( Q_1, Q_2, Q_3 \)

divides ranked scores into four equal parts

\[
\begin{align*}
\text{(minimum)} & \quad 25\% \\
Q_1 & \quad 25\% \\
Q_2 & \quad 25\% \\
Q_3 & \quad 25\% \\
\text{(maximum)} & \quad 25\%
\end{align*}
\]

**Percentiles**

Just as there are quartiles separating data into four parts, there are 99 percentiles denoted \( P_1, P_2, \ldots, P_{99} \), which partition the data into 100 groups.

**Finding the Percentile of a Given Score**

Percentile of value \( x \) =

\[
\frac{\text{number of values less than } x}{\text{total number of values}} \times 100
\]

**Converting from the \( k \)th Percentile to the Corresponding Data Value**

Notation

\[
L = \frac{k}{100} \cdot n
\]

- \( n \): total number of values in the data set
- \( k \): percentile being used
- \( L \): locator that gives the position of a value \( P_k \) in the data set

**Some Other Statistics**

- Interquartile Range (or IQR): \( Q_3 - Q_1 \)
- Semi-interquartile Range: \( \frac{Q_3 - Q_1}{2} \)
- Midquartile: \( \frac{Q_3 + Q_1}{2} \)
- 10 - 90 Percentile Range: \( P_{90} - P_{10} \)
Definition

- **Exploratory Data Analysis** is the process of using statistical tools (such as graphs, measures of center, and measures of variation) to investigate data sets in order to understand their important characteristics.

Definition

- An outlier is a value that is located very far away from almost all the other values.

Important Principles

- An outlier can have a dramatic effect on the mean.
- An outlier have a dramatic effect on the standard deviation.
- An outlier can have a dramatic effect on the scale of the histogram so that the true nature of the distribution is totally obscured.

Definitions

- For a set of data, the 5-number summary consists of the minimum value; the first quartile $Q_1$; the median (or second quartile $Q_2$); the third quartile, $Q_3$; and the maximum value.
- A boxplot (or box-and-whisker-diagram) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, $Q_1$; the median; and the third quartile, $Q_3$.

Boxplots

- Figure 2-16
- Figure 2-17

Definitions