Organizing Categorical Data: Summary Table

- A summary table indicates the frequency, amount, or percentage of items in a set of categories so that you can see differences between categories.

<table>
<thead>
<tr>
<th>How do you spend the holidays?</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>At home with family</td>
<td>45%</td>
</tr>
<tr>
<td>Travel to visit family</td>
<td>38%</td>
</tr>
<tr>
<td>Vacation</td>
<td>5%</td>
</tr>
<tr>
<td>Catching up on work</td>
<td>5%</td>
</tr>
<tr>
<td>Other</td>
<td>7%</td>
</tr>
</tbody>
</table>
Organizing Categorical Data: Bar Chart

- In a **bar chart**, a bar shows each category, the length of which represents the amount, frequency or percentage of values falling into a category.

![Bar Chart Example](image1)

Organizing Categorical Data: Pie Chart

- The **pie chart** is a circle broken up into slices that represent categories. The size of each slice of the pie varies according to the percentage in each category.

![Pie Chart Example](image2)
Organizing Categorical Data: Pareto Diagram

- Used to portray categorical data
- A bar chart, where categories are shown in descending order of frequency
- A cumulative polygon is shown in the same graph
- Used to separate the “vital few” from the “trivial many”
Class Exercise 1

2.1 A categorical variable has three categories with the following frequencies of occurrence:

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>13</td>
</tr>
<tr>
<td>B</td>
<td>28</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
</tr>
</tbody>
</table>

a. Compute the percentage of values in each category.
b. Construct a bar chart.
c. Construct a pie chart.
d. Construct a Pareto diagram.

Class Exercise 2

2.3 A survey of 705 workers asked how much they used the Internet at work. The results (USA Today Snapshots, March 21, 2006) were as follows:

<table>
<thead>
<tr>
<th>Use of the Internet at Work</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Too much</td>
<td>5</td>
</tr>
<tr>
<td>More than I should</td>
<td>4</td>
</tr>
<tr>
<td>Within limits</td>
<td>60</td>
</tr>
<tr>
<td>Very little</td>
<td>5</td>
</tr>
<tr>
<td>Do not use</td>
<td>26</td>
</tr>
</tbody>
</table>

a. Construct a bar chart, a pie chart, and a Pareto diagram.
b. Which graphical method do you think is best to portray these data?
c. Based on this survey, what conclusions can you reach about the use of the Internet at work?
Organizing Numerical Data: Frequency Distribution

- The **frequency distribution** is a summary table in which the data are arranged into numerically ordered class groupings.

- You must give attention to selecting the appropriate **number of class groupings** for the table, determining a suitable **width** of a class grouping, and establishing the **boundaries** of each class grouping to avoid overlapping.

- To determine the **width of a class interval**, you divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.

---

Organizing Numerical Data: Frequency Distribution Example

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27
Organizing Numerical Data: Frequency Distribution Example

- Sort raw data in ascending order:
  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find range: $58 - 12 = 46$
- Select number of classes: 5 (usually between 5 and 15)
- Compute class interval (width): 10 (46/5 then round up)
- Determine class boundaries (limits): 10, 20, 30, 40, 50, 60
- Compute class midpoints: 15, 25, 35, 45, 55
- Count observations & assign to classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 but less than 20</td>
<td>3</td>
<td>.15</td>
<td>15</td>
</tr>
<tr>
<td>20 but less than 30</td>
<td>6</td>
<td>.30</td>
<td>30</td>
</tr>
<tr>
<td>30 but less than 40</td>
<td>5</td>
<td>.25</td>
<td>25</td>
</tr>
<tr>
<td>40 but less than 50</td>
<td>4</td>
<td>.20</td>
<td>20</td>
</tr>
<tr>
<td>50 but less than 60</td>
<td>2</td>
<td>.10</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
<td><strong>1.00</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>
Organizing Numerical Data: The Histogram

- A graph of the data in a frequency distribution is called a **histogram**.

- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.

- The vertical axis is either **frequency**, **relative frequency**, or **percentage**.

- Bars of the appropriate heights are used to represent the number of observations within each class.

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 but less than 20</td>
<td>3</td>
<td>.15</td>
<td>15</td>
</tr>
<tr>
<td>20 but less than 30</td>
<td>6</td>
<td>.30</td>
<td>30</td>
</tr>
<tr>
<td>30 but less than 40</td>
<td>5</td>
<td>.25</td>
<td>25</td>
</tr>
<tr>
<td>40 but less than 50</td>
<td>4</td>
<td>.20</td>
<td>20</td>
</tr>
<tr>
<td>50 but less than 60</td>
<td>2</td>
<td>.10</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
<td><strong>1.00</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

**Histogram: Daily High Temperature**
Organizing Numerical Data: The Polygon

- A **percentage polygon** is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages.

- The **cumulative percentage polygon**, or **ogive**, displays the variable of interest along the X axis, and the cumulative percentages along the Y axis.

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 but less than 20</td>
<td>3</td>
<td>.15</td>
<td>15</td>
</tr>
<tr>
<td>20 but less than 30</td>
<td>6</td>
<td>.30</td>
<td>30</td>
</tr>
<tr>
<td>30 but less than 40</td>
<td>5</td>
<td>.25</td>
<td>25</td>
</tr>
<tr>
<td>40 but less than 50</td>
<td>4</td>
<td>.20</td>
<td>20</td>
</tr>
<tr>
<td>50 but less than 60</td>
<td>2</td>
<td>.10</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
<td><strong>1.00</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

(In a percentage polygon the vertical axis would be defined to show the percentage of observations per class)
Organizing Numerical Data: The Cumulative Percentage Polygon

<table>
<thead>
<tr>
<th>Class</th>
<th>Lower Boundary</th>
<th>% Less Than Lower Boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-20</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>20-30</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>30-40</td>
<td>30</td>
<td>45</td>
</tr>
<tr>
<td>40-50</td>
<td>40</td>
<td>70</td>
</tr>
<tr>
<td>50-60</td>
<td>50</td>
<td>90</td>
</tr>
</tbody>
</table>

Exercise 3

- Organize the above data into suitable class intervals
- Draw a histogram, frequency polygon and a cumulative frequency polygon

<table>
<thead>
<tr>
<th>18</th>
<th>60</th>
<th>3</th>
<th>89</th>
<th>63</th>
<th>52</th>
<th>54</th>
<th>30</th>
<th>25</th>
<th>83</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>24</td>
<td>94</td>
<td>7</td>
<td>95</td>
<td>51</td>
<td>44</td>
<td>84</td>
<td>20</td>
<td>38</td>
</tr>
<tr>
<td>24</td>
<td>75</td>
<td>86</td>
<td>64</td>
<td>63</td>
<td>9</td>
<td>70</td>
<td>97</td>
<td>86</td>
<td>29</td>
</tr>
<tr>
<td>17</td>
<td>84</td>
<td>28</td>
<td>43</td>
<td>100</td>
<td>63</td>
<td>72</td>
<td>29</td>
<td>59</td>
<td>90</td>
</tr>
<tr>
<td>75</td>
<td>36</td>
<td>72</td>
<td>9</td>
<td>86</td>
<td>44</td>
<td>70</td>
<td>81</td>
<td>3</td>
<td>34</td>
</tr>
</tbody>
</table>
Exercise 4

The following data (contained in the file
represent the cost of electricity during July 2006 for a random sample of 50 one-bedroom
apartments in a large city.

Raw Data on Utility Charges ($)
96 171 202 178 147 102 153 197 127 82
157 185 90 116 172 111 148 213 130 165
141 149 206 175 123 128 144 168 109 167
95 163 150 154 130 143 187 166 139 149
108 119 183 151 114 135 191 137 129 158

a. Form a frequency distribution and a percentage distribution that have class intervals with the upper class limits $99, $119, and so on.
b. Construct a histogram and a percentage polygon.
c. Form a cumulative percentage distribution and plot a cumulative percentage polygon.
d. Around what amount does the monthly electricity cost seem to be concentrated?

Principles of Excellent Graphs

- The graph should not distort the data.
- The graph should not contain unnecessary adornments (sometimes referred to as chart junk).
- The scale on the vertical axis should begin at zero.
- All axes should be properly labeled.
- The graph should contain a title.
- The simplest possible graph should be used for a given set of data.
Graphical Errors: Chart Junk

**Bad Presentation**

Minimum Wage

- 1960: $1.00
- 1970: $1.60
- 1980: $3.10
- 1990: $3.80

**Good Presentation**

Minimum Wage

$0 1 2 3 4

Graphical Errors: No Relative Basis

**Bad Presentation**

A's received by students.

<table>
<thead>
<tr>
<th>Freq.</th>
<th>FR</th>
<th>SO</th>
<th>JR</th>
<th>SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>200</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**Good Presentation**

A's received by students.

<table>
<thead>
<tr>
<th>%</th>
<th>FR</th>
<th>SO</th>
<th>JR</th>
<th>SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>20%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
</tr>
</tbody>
</table>

FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior
Graphical Errors: Compressing the Vertical Axis

Bad Presentation
Quarterly Sales

Good Presentation
Quarterly Sales

Graphical Errors: No Zero Point on the Vertical Axis

Bad Presentation
Monthly Sales

Good Presentations
Monthly Sales

Graphing the first six months of sales
Summary Definitions

- The **central tendency** is the extent to which all the data values group around a typical or central value.
- The **variation** is the amount of dispersion, or scattering, of values.
- The **shape** is the pattern of the distribution of values from the lowest value to the highest value.
Measures of Central Tendency
The Arithmetic Mean

- The arithmetic mean (mean) is the most common measure of central tendency

For a sample of size \( n \):

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}
\]

Sample size
Observed values

Measures of Central Tendency
The Arithmetic Mean

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)

\[
\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3
\]

\[
\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4
\]
Measures of Central Tendency

The Median

- In an ordered array, the median is the “middle” number (50% above, 50% below)

\[ \text{Median} = \frac{n + 1}{2} \]

- Not affected by extreme values

Measures of Central Tendency

Locating the Median

- The median of an ordered set of data is located at the \( \frac{n + 1}{2} \) ranked value.

- If the number of values is odd, the median is the middle number.
- If the number of values is even, the median is the average of the two middle numbers.
- Note that \( \frac{n + 1}{2} \) is NOT the value of the median, only the position of the median in the ranked data.
Measures of Central Tendency

The Mode

- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes

0   1   2   3   4   5   6   7   8   9   10   11   12   13   14

Mode = 9

Measures of Central Tendency

Review Example

House Prices:

<table>
<thead>
<tr>
<th>Price</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2,000,000</td>
<td>5</td>
</tr>
<tr>
<td>500,000</td>
<td>3</td>
</tr>
<tr>
<td>300,000</td>
<td>1</td>
</tr>
<tr>
<td>100,000</td>
<td>1</td>
</tr>
<tr>
<td>100,000</td>
<td>1</td>
</tr>
</tbody>
</table>

Sum 3,000,000

- **Mean:** \( \frac{3,000,000}{5} = 600,000 \)
- **Median:** middle value of ranked data = 300,000
- **Mode:** most frequent value = 100,000
Measures of Central Tendency
Which Measure to Choose?

- The **mean** is generally used, unless extreme values (outliers) exist.

- Then **median** is often used, since the median is not sensitive to extreme values. For example, median home prices may be reported for a region; it is less sensitive to outliers.
### Quartile Measures

- Quartiles split the ranked data into 4 segments with an equal number of values per segment.

<table>
<thead>
<tr>
<th>25%</th>
<th>25%</th>
<th>25%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q₁</td>
<td>Q₂</td>
<td>Q₃</td>
<td></td>
</tr>
</tbody>
</table>

- The first quartile, Q₁, is the value for which 25% of the observations are smaller and 75% are larger.
- Q₃ is the same as the median (50% are smaller, 50% are larger).
- Only 25% of the values are greater than the third quartile.

### Locating Quartiles

Find a quartile by determining the value in the appropriate position in the ranked data, where:

- First quartile position: \( Q_1 = \frac{(n+1)}{4} \) ranked value
- Second quartile position: \( Q_2 = \frac{(n+1)}{2} \) ranked value
- Third quartile position: \( Q_3 = \frac{3(n+1)}{4} \) ranked value

Where \( n \) is the number of observed values.
Quartile Measures

Guidelines

- Rule 1: If the result is a whole number, then the quartile is equal to that ranked value.

- Rule 2: If the result is a fraction half (2.5, 3.5, etc), then the quartile is equal to the average of the corresponding ranked values.

- Rule 3: If the result is neither a whole number or a fractional half, you round the result to the nearest integer and select that ranked value.

Locating the First Quartile

- Example: Find the first quartile

Sample Data in Ordered Array: 11, 12, 13, 16, 16, 17, 18, 21, 22

First, note that \( n = 9 \).

\( Q_1 = \) is in the \( \frac{9+1}{4} = 2.5 \) ranked value of the ranked data, so use the value half way between the 2\(^{nd}\) and 3\(^{rd}\) ranked values,

so \( Q_1 = 12.5 \)

\( Q_1 \) and \( Q_3 \) are measures of non-central location
\( Q_2 \) = median, a measure of central tendency
Exercise 5

3.9 The data in the file bankcost.xls contain the bounced check fees, in dollars, for a sample of 23 banks for direct-deposit customers who maintain a $100 balance:
26 28 20 20 21 22 25 25 18 25 15 20
18 20 25 22 30 30 30 15 20 29


a. Compute the mean, median, first quartile, and third quartile.
b. Compute the variance, standard deviation, range, interquartile range, coefficient of variation, and Z scores.
c. Are the data skewed? If so, how?
d. Based on the results of (a) through (c), what conclusions can you reach concerning the bounced check fees?

Measures of Central Tendency
The Geometric Mean

- Geometric mean
  - Used to measure the rate of change of a variable over time
    \[ \bar{X}_G = \left( X_1 \times X_2 \times \cdots \times X_n \right)^{1/n} \]

- Geometric mean rate of return
  - Measures the status of an investment over time
    \[ \bar{R}_G = \left[ (1+R_1) \times (1+R_2) \times \cdots \times (1+R_n) \right]^{1/n} - 1 \]
  - Where \( R_i \) is the rate of return in time period \( i \)
Measures of Central Tendency
The Geometric Mean

An investment of $100,000 declined to $50,000 at the end of year one and rebounded to $100,000 at end of year two:

\[ X_1 = $100,000 \quad X_2 = $50,000 \quad X_3 = $100,000 \]

50% decrease \quad 100% increase

The overall two-year return is zero, since it started and ended at the same level.

Measures of Central Tendency
The Geometric Mean

Use the 1-year returns to compute the arithmetic mean and the geometric mean:

Arithmetic mean rate of return:

\[ \bar{X} = \frac{(-.5) + (1)}{2} = .25 \]

Misleading result

Geometric mean rate of return:

\[ \bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{1/n} - 1 \]

\[ = [1 + (-.5) \times (1 + (1))]^{1/2} - 1 \]

\[ = [.50 \times (2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\% \]

More accurate result
Measures of Central Tendency

**Central Tendency**

- **Arithmetic Mean**
  \[ \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \]
  Middle value in the ordered array

- **Median**

- **Mode**
  Most frequently observed value

- **Geometric Mean**
  \[ X_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n} \]

Measures of Variation

- **Variation** measures the spread, or dispersion, of values in a data set.
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation
Measures of Variation

Range

- Simplest measure of variation
- Difference between the largest and the smallest values:

\[ \text{Range} = X_{\text{largest}} - X_{\text{smallest}} \]

Example:

\[ \text{Range} = 13 - 1 = 12 \]

Disadvantages of the Range

- Ignores the way in which data are distributed
- Sensitive to outliers

\[ 1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,4,5 \]

\[ \text{Range} = 5 - 1 = 4 \]

\[ 1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,4,120 \]

\[ \text{Range} = 120 - 1 = 119 \]
Measures of Variation
Interquartile Range

- Problems caused by outliers can be eliminated by using the **interquartile range**.

- The IQR can eliminate some high and low values and calculate the range from the remaining values.

- Interquartile range = 3rd quartile – 1st quartile
  \[ = Q_3 - Q_1 \]

Example:

<table>
<thead>
<tr>
<th>X minimum</th>
<th>12</th>
<th>Q_1</th>
<th>30</th>
<th>Median (Q_2)</th>
<th>45</th>
<th>Q_3</th>
<th>57</th>
<th>X maximum</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td></td>
<td></td>
<td>25%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Interquartile range = 57 – 30 = 27
Measures of Variation

Variance

- The **variance** is the average (approximately) of squared deviations of values from the mean.

Sample variance:

\[ S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1} \]

Where
- \( \bar{X} \) = arithmetic mean
- \( n \) = sample size
- \( X_i \) = i\text{th} value of the variable X

Measures of Variation

Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

Sample standard deviation:

\[ S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}} \]
Measures of Variation
Standard Deviation

Steps for Computing Standard Deviation

1. Compute the difference between each value and the mean.
2. Square each difference.
3. Add the squared differences.
4. Divide this total by n-1 to get the sample variance.
5. Take the square root of the sample variance to get the sample standard deviation.

Sample Data \( (X_i) \): 10 12 14 15 17 18 18 24

\[ n = 8 \quad \text{Mean} = \bar{X} = 16 \]

\[ S = \sqrt{\frac{(10 - \bar{X})^2 + (12 - \bar{X})^2 + (14 - \bar{X})^2 + \cdots + (24 - \bar{X})^2}{n - 1}} \]

\[ = \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \cdots + (24 - 16)^2}{8 - 1}} \]

\[ = \sqrt{\frac{126}{7}} = 4.2426 \quad \text{A measure of the “average” scatter around the mean} \]
Measures of Variation
Comparing Standard Deviation

Data A
Mean = 15.5
S = 3.338

Data B
Mean = 15.5
S = 0.926

Data C
Mean = 15.5
S = 4.570

Exercise 5

3.9 The data in the file bankcost.xls contain the bounced check fees, in dollars, for a sample of 23 banks for direct-deposit customers who maintain a $100 balance.

26 28 20 20 21 22 25 25 18 25 15 20
18 20 25 22 30 30 30 15 20 29


a. Compute the mean, median, first quartile, and third quartile.
b. Compute the variance, standard deviation, range, interquartile range, coefficient of variation, and Z scores.
c. Are the data skewed? If so, how?
d. Based on the results of (a) through (c), what conclusions can you reach concerning the bounced check fees?
Measures of Variation
Comparing Standard Deviation

Small standard deviation

Large standard deviation

Measures of Variation
Summary Characteristics

- The more the data are spread out, the greater the range, interquartile range, variance, and standard deviation.
- The more the data are concentrated, the smaller the range, interquartile range, variance, and standard deviation.
- If the values are all the same (no variation), all these measures will be zero.
- None of these measures are ever negative.
Coefficient of Variation

- The coefficient of variation is the standard deviation divided by the mean, multiplied by 100.
- It is always expressed as a percentage. (%) 
- It shows variation relative to mean.
- The CV can be used to compare two or more sets of data measured in different units.

\[ CV = \left( \frac{S}{X} \right) \cdot 100\% \]

Stock A:
- Average price last year = $50
- Standard deviation = $5

\[ CV_A = \left( \frac{S}{X} \right) \cdot 100\% = \frac{5}{50} \cdot 100\% = 10\% \]

Stock B:
- Average price last year = $100
- Standard deviation = $5

\[ CV_B = \left( \frac{S}{X} \right) \cdot 100\% = \frac{5}{100} \cdot 100\% = 5\% \]

Both stocks have the same standard deviation, but stock B is less variable relative to its price.
Exercise 5

3.9 The data in the file bankcost.xls contain the bounced check fees, in dollars, for a sample of 23 banks for direct-deposit customers who maintain a $100 balance:

26 28 20 20 21 22 25 25 18 25 15 20
18 20 25 25 22 30 30 30 15 20 29


a. Compute the mean, median, first quartile, and third quartile.
b. Compute the variance, standard deviation, range, interquartile range, coefficient of variation, and Z scores.
c. Are the data skewed? If so, how?
d. Based on the results of (a) through (c), what conclusions can you reach concerning the bounced check fees?

Exercise 6

3.11 The data in the file chicken.xls contain the total fat, in grams, per serving, for a sample of 20 chicken sandwiches from fast-food chains. The data is as follows:

7 8 4 5 16 20 20 24 19 30
23 30 25 19 29 29 30 30 40 56


a. Compute the mean, median, first quartile, and third quartile.
b. Compute the variance, standard deviation, range, interquartile range, coefficient of variation, and Z scores.
c. Are there any outliers? Explain.
d. Based on the results of (a) through (c), what conclusions can you reach concerning the total fat of chicken sandwiches?
Numerical Descriptive Measures for a Population

- Descriptive statistics discussed previously described a sample, not the population.

- Summary measures describing a population, called parameters, are denoted with Greek letters.

- Important population parameters are the population mean, variance, and standard deviation.

Population Mean

- The population mean is the sum of the values in the population divided by the population size, N.

\[
\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}
\]

Where
- \( \mu \) = population mean
- \( N \) = population size
- \( X_i \) = \( i \)th value of the variable X
Population Variance

- The population variance is the average of squared deviations of values from the mean.

\[
\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}
\]

Where:
- \( \mu \) = population mean
- \( N \) = population size
- \( X_i \) = \( i \)th value of the variable \( X \)

Population Standard Deviation

- The population standard deviation is the most commonly used measure of variation.
- It has the same units as the original data.

\[
\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}
\]

Where:
- \( \mu \) = population mean
- \( N \) = population size
- \( X_i \) = \( i \)th value of the variable \( X \)
Sample statistics versus population parameters

<table>
<thead>
<tr>
<th>Measure</th>
<th>Population Parameter</th>
<th>Sample Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>( \mu )</td>
<td>( \bar{X} )</td>
</tr>
<tr>
<td>Variance</td>
<td>( \sigma^2 )</td>
<td>( S^2 )</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>( \sigma )</td>
<td>( S )</td>
</tr>
</tbody>
</table>

The Empirical Rule

- The **empirical rule** approximates the variation of data in bell-shaped distributions.

Approximately 68% of the data in a bell-shaped distribution lies within one standard deviation of the mean, or \( \pm 1\sigma \)

\[ 68\% \]

\[ \pm 1\sigma \]
The Empirical Rule

- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviation of the mean, or $\pm 2\sigma$
- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviation of the mean, or $\pm 3\sigma$

Exploratory Data Analysis

The Five Number Summary

- The five numbers that describe the spread of data are:
  - Minimum
  - First Quartile ($Q_1$)
  - Median ($Q_2$)
  - Third Quartile ($Q_3$)
  - Maximum
Pitfalls in Numerical Descriptive Measures

- Data analysis is **objective**
  - Analysis should report the summary measures that best meet the assumptions about the data set.

- Data interpretation is **subjective**
  - Interpretation should be done in fair, neutral and clear manner.

Ethical Considerations

Numerical descriptive measures:

- Should document both good and bad results
- Should be presented in a fair, objective and neutral manner
- Should not use inappropriate summary measures to distort facts