Measures of Central Tendency: Ungrouped Data

- Measures of central tendency yield information about “particular places or locations in a group of numbers.”
- Common Measures of Location
  - Mode
  - Median
  - Mean
  - Percentiles
  - Quartiles

Mode

- The most frequently occurring value in a data set
- Applicable to all levels of data measurement (nominal, ordinal, interval, and ratio)
- Bimodal -- Data sets that have two modes
- Multimodal -- Data sets that contain more than two modes

Mode -- Example

- The mode is 44.
- There are more 44s than any other value.

Median

- Middle value in an ordered array of numbers.
- Applicable for ordinal, interval, and ratio data
- Not applicable for nominal data
- Unaffected by extremely large and extremely small values.

Median: Computational Procedure

- First Procedure
  - Arrange the observations in an ordered array.
  - If there is an odd number of terms, the median is the middle term of the ordered array.
  - If there is an even number of terms, the median is the average of the middle two terms.
- Second Procedure
  - The median’s position in an ordered array is given by (n+1)/2.

Median: Example with an Odd Number of Terms

Ordered Array
3 4 5 7 8 9 11 14 15 16 17 19 19 20 21 22

- There are 17 terms in the ordered array.
- Position of median = (n+1)/2 = (17+1)/2 = 9
- The median is the 9th term, 15.
- If the 22 is replaced by 100, the median is 15.
- If the 3 is replaced by -103, the median is 15.
Median: Example with an Even Number of Terms

Ordered Array
3 4 5 7 8 9 11 14 15 16 17 19 20 21

- There are 16 terms in the ordered array.
- Position of median = \( (n+1)/2 = (16+1)/2 = 8.5 \)
- The median is between the 8th and 9th terms, 14.5.
- If the 21 is replaced by 100, the median is 14.5.
- If the 3 is replaced by -88, the median is 14.5.

Arithmetic Mean

- Commonly called ‘the mean’
- is the average of a group of numbers
- Applicable for interval and ratio data
- Not applicable for nominal or ordinal data
- Affected by each value in the data set, including extreme values
- Computed by summing all values in the data set and dividing the sum by the number of values in the data set

Population Mean

\[
\mu = \frac{\sum X}{N} = \frac{24 + 13 + 19 + 26 + 11}{5} = \frac{93}{5} = 18.6
\]

Sample Mean

\[
\bar{X} = \frac{\sum X}{n} = \frac{57 + 86 + 42 + 38 + 90 + 66}{6} = \frac{379}{6} = 63.167
\]

Percentiles

- Measures of central tendency that divide a group of data into 100 parts
- At least \( n\% \) of the data lie below the \( n^{th} \) percentile, and at most \( (100 - n)\% \) of the data lie above the \( n^{th} \) percentile
- Example: 90th percentile indicates that at least 90% of the data lie below it, and at most 10% of the data lie above it
- The median and the 50th percentile are the same.
- Applicable for ordinal, interval, and ratio data
- Not applicable for nominal data

Percentiles: Computational Procedure

- Organize the data into an ascending ordered array.
- Calculate the percentile location: \( i = \frac{P}{100} (n) \)
- Determine the percentile’s location and its value.
- If \( i \) is a whole number, the percentile is the average of the values at the \( i \) and \( (i+1) \) positions.
- If \( i \) is not a whole number, the percentile is at the \( (i+1) \) position in the ordered array.
**Percentiles: Example**

- Raw Data: 14, 12, 19, 23, 5, 13, 28, 17
- Ordered Array: 5, 12, 13, 14, 17, 19, 23, 28
- Location of 30th percentile:
  \[ i = \frac{30}{100} \times (8) = 2.4 \]
  - The location index, \( i \), is not a whole number; \( i+1 = 2.4+1 = 3.4 \); the whole number portion is 3; the 30th percentile is at the 3rd location of the array; the 30th percentile is 13.

**Quartiles**

- Measures of central tendency that divide a group of data into four subgroups
- \( Q_1 \): 25% of the data set is below the first quartile
- \( Q_2 \): 50% of the data set is below the second quartile
- \( Q_3 \): 75% of the data set is below the third quartile
- \( Q_1 \) is equal to the 25th percentile
- \( Q_2 \) is located at the 50th percentile and equals the median
- \( Q_3 \) is equal to the 75th percentile
- Quartile values are not necessarily members of the data set

**Quartiles: Example**

- Ordered array: 106, 109, 114, 116, 121, 122, 125, 129
- \( Q_1 \):
  \[ i = \frac{25}{100} \times (8) = 2 \quad Q = \frac{109+114}{2} = 111.5 \]
- \( Q_2 \):
  \[ i = \frac{50}{100} \times (8) = 4 \quad Q = \frac{116+121}{2} = 118.5 \]
- \( Q_3 \):
  \[ i = \frac{75}{100} \times (8) = 6 \quad Q = \frac{122+125}{2} = 123.5 \]

**Variability**

- No Variability in Cash Flow
- Variability in Cash Flow
Measures of Variability: Ungrouped Data

- Measures of variability describe the spread or the dispersion of a set of data.
- Common Measures of Variability
  - Range
  - Interquartile Range
  - Mean Absolute Deviation
  - Variance
  - Standard Deviation
  - Z scores
  - Coefficient of Variation

Range

- The difference between the largest and the smallest values in a set of data
- Simple to compute
- Ignores all data points except two extremes
- Example:
  - Range
  - Largest - Smallest
  - 48 - 35 = 13

Interquartile Range

- Range of values between the first and third quartiles
- Range of the “middle half”
- Less influenced by extremes

Mean Absolute Deviation

- Average of the absolute deviations from the mean

Population Variance

- Average of the squared deviations from the arithmetic mean

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

- **Square root of the variance**

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

<table>
<thead>
<tr>
<th>(X)</th>
<th>(X - \mu)</th>
<th>((X - \mu)^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-8</td>
<td>64</td>
</tr>
<tr>
<td>9</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>16</td>
<td>+3</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>+4</td>
<td>16</td>
</tr>
<tr>
<td>18</td>
<td>+5</td>
<td>25</td>
</tr>
</tbody>
</table>

\[
\sigma = \sqrt{\frac{130}{5}} = 26.0
\]

\[
\sigma = 5.1
\]

### Sample Variance

- **Average of the squared deviations from the arithmetic mean**

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

<table>
<thead>
<tr>
<th>(X)</th>
<th>(X - \bar{X})</th>
<th>((X - \bar{X})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,398</td>
<td>625</td>
<td>390,625</td>
</tr>
<tr>
<td>1,844</td>
<td>71</td>
<td>5,041</td>
</tr>
<tr>
<td>1,539</td>
<td>-234</td>
<td>54,756</td>
</tr>
<tr>
<td>1,311</td>
<td>-462</td>
<td>213,444</td>
</tr>
<tr>
<td>7,092</td>
<td>-6</td>
<td>663,866</td>
</tr>
</tbody>
</table>

\[
S^2 = \frac{663,866}{3} = 221,288.67
\]

### Sample Standard Deviation

- **Square root of the sample variance**

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

<table>
<thead>
<tr>
<th>(X)</th>
<th>(X - X)</th>
<th>((X - X)^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,398</td>
<td>625</td>
<td>390,625</td>
</tr>
<tr>
<td>1,844</td>
<td>71</td>
<td>5,041</td>
</tr>
<tr>
<td>1,539</td>
<td>-234</td>
<td>54,756</td>
</tr>
<tr>
<td>1,311</td>
<td>-462</td>
<td>213,444</td>
</tr>
<tr>
<td>7,092</td>
<td>-6</td>
<td>663,866</td>
</tr>
</tbody>
</table>

\[
S = \sqrt{\frac{663,866}{3}} = 470.41
\]

### Uses of Standard Deviation

- **Indicator of financial risk**
- **Quality Control**
  - Construction of quality control charts
  - Process capability studies
- **Comparing populations**
  - Household incomes in two cities
  - Employee absenteeism at two companies

### Standard Deviation as an Indicator of Financial Risk

<table>
<thead>
<tr>
<th>Financial Security</th>
<th>Annualized Rate of Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15%</td>
</tr>
<tr>
<td>B</td>
<td>15%</td>
</tr>
</tbody>
</table>

### Empirical Rule

- **Data are normally distributed (or approximately normal)**

\[
\mu \pm 1\sigma = 68\%
\]

\[
\mu \pm 2\sigma = 95\%
\]

\[
\mu \pm 3\sigma = 99.7\%
\]
Chebyshev’s Theorem

- Applies to all distributions

\[ P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2} \]
for \( k > 1 \)

<table>
<thead>
<tr>
<th>Number of Standard Deviations</th>
<th>Distance from the Mean</th>
<th>Minimum Proportion of Values Falling Within Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K = 2 )</td>
<td>( \mu \pm 2\sigma )</td>
<td>1 - ( \frac{1}{2^2} = 0.75 )</td>
</tr>
<tr>
<td>( K = 3 )</td>
<td>( \mu \pm 3\sigma )</td>
<td>1 - ( \frac{1}{3^2} = 0.89 )</td>
</tr>
<tr>
<td>( K = 4 )</td>
<td>( \mu \pm 4\sigma )</td>
<td>1 - ( \frac{1}{4^2} = 0.94 )</td>
</tr>
</tbody>
</table>

Coefficient of Variation

- Ratio of the standard deviation to the mean, expressed as a percentage
- Measurement of relative dispersion

\[ C.V. = \frac{\sigma}{\mu} \times 100 \]

\[ \mu_1 = 29 \]
\[ \sigma_1 = 4.6 \]
\[ C.V._1 = \frac{\sigma_1}{\mu_1} \times 100 = \frac{4.6}{29} \times 100 = 15.86 \%
\]

\[ \mu_2 = 84 \]
\[ \sigma_2 = 10 \]
\[ C.V._2 = \frac{\sigma_2}{\mu_2} \times 100 = \frac{10}{84} \times 100 = 11.90 \%
\]

Measures of Central Tendency and Variability: Grouped Data

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
- Measures of Variability
  - Variance
  - Standard Deviation
  - Mean Absolute Deviation

Mean of Grouped Data

- Weighted average of class midpoints
- Class frequencies are the weights

\[ \mu = \frac{\sum f M}{\sum f} \]
\[ = \frac{\sum f M}{N} \]
\[ = \frac{f_1 M_1 + f_2 M_2 + \ldots + f_n M_n}{f_1 + f_2 + \ldots + f_n} \]
Calculation of Grouped Mean

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
<th>Class Midpoint</th>
<th>fM</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-under 30</td>
<td>6</td>
<td>25</td>
<td>150</td>
</tr>
<tr>
<td>30-under 40</td>
<td>18</td>
<td>35</td>
<td>630</td>
</tr>
<tr>
<td>40-under 50</td>
<td>11</td>
<td>45</td>
<td>495</td>
</tr>
<tr>
<td>50-under 60</td>
<td>11</td>
<td>55</td>
<td>605</td>
</tr>
<tr>
<td>60-under 70</td>
<td>3</td>
<td>65</td>
<td>195</td>
</tr>
<tr>
<td>70-under 80</td>
<td>1</td>
<td>75</td>
<td>75</td>
</tr>
</tbody>
</table>

\[ \mu = \frac{\sum fm}{\sum f} = \frac{2150}{50} = 43.0 \]

Median of Grouped Data

\[ \text{Median} = L + \frac{\frac{N}{2} - cf}{f_{\text{med}}} (W) \]

Where:
- \( L \) = the lower limit of the median class
- \( cf \) = cumulative frequency of class preceding the median class
- \( f_{\text{med}} \) = frequency of the median class
- \( W \) = width of the median class
- \( N \) = total of frequencies

Mode of Grouped Data

- Midpoint of the modal class
- Modal class has the greatest frequency

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-under 40</td>
<td>18</td>
<td>30-40</td>
</tr>
<tr>
<td>40-under 50</td>
<td>11</td>
<td>40-50</td>
</tr>
<tr>
<td>50-under 60</td>
<td>11</td>
<td>50-60</td>
</tr>
<tr>
<td>60-under 70</td>
<td>3</td>
<td>60-70</td>
</tr>
<tr>
<td>70-under 80</td>
<td>1</td>
<td>70-80</td>
</tr>
</tbody>
</table>

Variance and Standard Deviation of Grouped Data

Population

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

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

Sample

\[ S^2 = \frac{\sum f(M - \overline{X})^2}{n-1} \]

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

Population Variance and Standard Deviation of Grouped Data

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>f</th>
<th>M</th>
<th>fM</th>
<th>(M-\mu)²</th>
<th>(M-\mu)² \cdot f</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-under 30</td>
<td>6</td>
<td>25</td>
<td>150</td>
<td>18</td>
<td>324</td>
</tr>
<tr>
<td>30-under 40</td>
<td>18</td>
<td>35</td>
<td>630</td>
<td>8</td>
<td>64</td>
</tr>
<tr>
<td>40-under 50</td>
<td>11</td>
<td>55</td>
<td>605</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>50-under 60</td>
<td>11</td>
<td>65</td>
<td>195</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>60-under 70</td>
<td>3</td>
<td>75</td>
<td>215</td>
<td>12</td>
<td>144</td>
</tr>
<tr>
<td>70-under 80</td>
<td>1</td>
<td>75</td>
<td>720</td>
<td>22</td>
<td>484</td>
</tr>
</tbody>
</table>

\[ \sigma = \sqrt{\frac{\sum f(M - \mu)^2}{N}} = \sqrt{\frac{7200}{50}} = 144 \]

\[ \sigma = \sqrt{144} = 12 \]
Measures of Shape

- **Skewness**
  - Absence of symmetry
  - Extreme values in one side of a distribution
- **Kurtosis**
  - peakedness of a distribution
  - Leptokurtic: high and thin
  - Mesokurtic: normal shape
  - Platykurtic: flat and spread out
- **Box and Whisker Plots**
  - Graphic display of a distribution
  - Reveals skewness

Skewness

- **Negatively (Left) Skewed**
- **Symmetric (Not Skewed)**
- **Positively (Right) Skewed**

Coefficient of Skewness

- Summary measure for skewness
  - If $S < 0$, the distribution is **negatively skewed** (skewed to the left).
  - If $S = 0$, the distribution is **symmetric** (not skewed).
  - If $S > 0$, the distribution is **positively skewed** (skewed to the right).

**Skewness & Kurtosis**

- What do we mean by symmetry and positive and negative skewness? Kurtosis? Properties??
  \[
  \text{Skewness} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(N-1)SD^3} ; \quad \text{Kurtosis} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{(N-1)SD^4}
  \]

- Skewness in linearly invariant
  \[\text{Sk}(aX+b)=\text{Sk}(X)\]

- Skewness is a measure of unsymmetry

- Kurtosis is a measure of flatness

- Both are used to quantify departures from StdNormal

- Skewness(StdNorm)=0; Kurtosis(StdNorm)=3

Kurtosis

- Peakedness of a distribution
  - Leptokurtic: high and thin
  - Mesokurtic: normal in shape
  - Platykurtic: flat and spread out
Box and Whisker Plot

- Five specific values are used:
  - Median, \( Q_2 \)
  - First quartile, \( Q_1 \)
  - Third quartile, \( Q_3 \)
  - Minimum value in the data set
  - Maximum value in the data set
- Inner Fences
  - \( IQR = Q_3 - Q_1 \)
  - Lower inner fence = \( Q_1 - 1.5 \times IQR \)
  - Upper inner fence = \( Q_3 + 1.5 \times IQR \)
- Outer Fences
  - Lower outer fence = \( Q_1 - 3.0 \times IQR \)
  - Upper outer fence = \( Q_3 + 3.0 \times IQR \)

Skewness: Box and Whisker Plots, and Coefficient of Skewness

- Negatively Skewed: \( S < 0 \)
- Symmetric (Not Skewed): \( S = 0 \)
- Positively Skewed: \( S > 0 \)

Pearson Product-Moment Correlation Coefficient

\[
r = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}
\]

Three Degrees of Correlation

- \( r < 0 \)
- \( r > 0 \)
- \( r = 0 \)

Computation of \( r \) for the Economics Example (Part 1)

<table>
<thead>
<tr>
<th>No</th>
<th>Day</th>
<th>Interest</th>
<th>Futures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>2</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>5</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>6</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>7</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>8</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>9</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>10</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>11</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
<tr>
<td>12</td>
<td>0.03</td>
<td>233</td>
<td>64.481</td>
</tr>
</tbody>
</table>

Summations:
- \( SS_Y = 729.220 \)
- \( SS_X = 619.207 \)
- \( SS_{XY} = 2135.087 \)
Computation of $r$ for the Economics Example (Part 2)

\[ r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right) \left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right)}} \]

\[ = \frac{2115(2725) - (2115)^2}{\sqrt{(220)(225) - (2115)^2} \sqrt{(220)(225) - (2725)^2}} \]

\[ = 0.815 \]

Scatter Plot and Correlation Matrix for the Economics Example

Palindrome Testing Program (1 of 5)

// test for palindrome property – Cstrings vs Strings!
#include <iostream>
#include <string>
#include <cctype>
using namespace std;

void swap(char& lhs, char& rhs);
//swaps char args corresponding to parameters lhs and rhs

string reverse(const string& str);
//returns a copy of arg corresponding to parameter
//str with characters in reverse order.

string removePunct(const string& src, const string& punct);
//returns copy of string src with characters
//in string punct removed.

string makeLower (const string& s);
//returns a copy of parameter s that has all upper case
//characters forced to lower case, other characters unchanged.
//Uses <cctype>, which provides tolower

bool isPal(const string& this_String);
//uses makeLower, removePunct.
//if  this_String is a palindrome,
//    return true;
//else
//    return false;

//  Palindrome Testing Program (2 of 5)
int main()
{
    string str;
    cout << "Enter a candidate for palindrome test 
followed by pressing return."
    getline(cin, str);
    if (isPal(str))
        cout << """ << str + "" is a palindrome 
"" << endl;
    else
        cout << """ << str + "" is not a palindrome 
"" << endl;
    return 0;
}

void swap(char& lhs, char& rhs)
{
    char tmp = lhs;
    lhs = rhs;
    rhs = tmp;
}

//  Palindrome Testing Program (3 of 5)
string reverse(const string& str)
{
    int start = 0;
    int end = str.length();
    string temp(s);
    while (start < end)
    {
        end--;
        swap(temp[start], temp[end]);
        start++;
    }
    return temp;
}

//returns a copy of src with characters in punct removed
string removePunct(const string& src, const string& punct)
{
    string no_punct;
    int src_len = src.length();
    int punct_len = punct.length();
    for(int i = 0; i < src_len; i++)
    {
        string aChar = src.substr(i,1);
        int location = punct.find(aChar, 0);
        //find location of successive characters
        //not in punct
        if (location < 0 || location >= punct_len)
            no_punct = no_punct + aChar; //aChar not in punct -- keep it
    }
    return no_punct;
}

//  Palindrome Testing Program (4 of 5)
string reverse(const string& str)
{
    int start = 0;
    int end = str.length();
    string temp(s);
    while (start < end)
    {
        end--;
        swap(temp[start], temp[end]);
        start++;
    }
    return temp;
}
// Palindrome Testing Program (5 of 5)
// Uses functions makeLower, removePunct.
// Returned value:
//  If this_String is a palindrome,
//   return true;
// else
//   return false;

bool isPal(const string& this_String) {
    string punctuation(",:;?!
\" "); // includes a blank
    string str(this_String);
    str = makeLower(str);
    string lowerStr = removePunct(str, punctuation);
    return lowerStr == reverse(lowerStr);
}