Measures of central tendency for ungrouped data

- Graphs are very helpful to describe the basic shape of a data distribution; “a picture is worth a thousand words.” There are limitations, however to the use of graphs.
- One way to overcome graph problems is to use numerical measures, which can be calculated for either a sample or a population.
- Numerical descriptive measures associated with a population of measurements are called parameters; those computed from sample measurements are called statistics.
- A measure of central tendency gives the center of a histogram or frequency distribution curve.
- The measures are: mean, median, and mode.
Measures of central tendency for ungrouped data

- Data that give information on each member of the population or sample individually are called ungrouped data, whereas grouped data are presented in the form of a frequency distribution table.

**Mean**

- The mean (average) is the most frequently used measure of central tendency.
- The mean for ungrouped data is obtained by dividing the sum of all values by the number of values in the data set. Thus,

\[
\text{Mean for population: } \mu = \frac{\sum x}{N}
\]

\[
\text{Mean for sample: } \bar{x} = \frac{\sum x}{n}
\]

**Example:** The following table gives the 2002 total payrolls of five MLB teams.

<table>
<thead>
<tr>
<th>MLB team</th>
<th>2002 total payroll (millions of dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaheim Angels</td>
<td>62</td>
</tr>
<tr>
<td>Atlanta Braves</td>
<td>93</td>
</tr>
<tr>
<td>New York Yankees</td>
<td>126</td>
</tr>
<tr>
<td>St. Louis Cardinals</td>
<td>75</td>
</tr>
<tr>
<td>Tampa Bay Devil Rays</td>
<td>34</td>
</tr>
</tbody>
</table>

The Mean is a balancing point.
Sometimes a data set may contain a few very small or a few very large values. Such values are called outliers or extreme values.

We should be very cautious when using the mean. It may not always be the best measure of central tendency.

**Example:** The following table lists the 2000 populations (in thousands) of five Pacific states.

<table>
<thead>
<tr>
<th>State</th>
<th>Population (thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Washington</td>
<td>5894</td>
</tr>
<tr>
<td>Oregon</td>
<td>3421</td>
</tr>
<tr>
<td>Alaska</td>
<td>627</td>
</tr>
<tr>
<td>Hawaii</td>
<td>An outlier 1212</td>
</tr>
<tr>
<td>California</td>
<td>627</td>
</tr>
</tbody>
</table>

**Excluding California**

\[
\text{Mean} = \frac{5894 + 3421 + 627 + 1212}{4} = 2788.5
\]

**Including California**

\[
\text{Mean} = \frac{5894 + 3421 + 627 + 1212 + 33,872}{5} = 9005.2
\]

---

**Weighted Mean**

Sometimes we may assign weight (importance) to each observation before we calculate the mean.

A mean computed in this manner is referred to as a weighted mean and it is given by

\[
\bar{x} = \frac{\sum W_i x_i}{\sum W_i}
\]

where \(x_i\) is the value of observation \(i\) and \(W_i\) is weight for observation \(i\).
Example: Consider the following sample of four purchases of one stock in the KSE. Find the average cost of the stock.

<table>
<thead>
<tr>
<th>Purchase</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.300</td>
<td>5,000</td>
</tr>
<tr>
<td>2</td>
<td>.325</td>
<td>15,000</td>
</tr>
<tr>
<td>3</td>
<td>.350</td>
<td>10,000</td>
</tr>
<tr>
<td>4</td>
<td>.295</td>
<td>20,000</td>
</tr>
</tbody>
</table>

Median

The median is the value of the middle term in a data set has been ranked in increasing order.

\[
\text{Median } = \left( \frac{n+1}{2} \right) \text{ value in a ranked data set}
\]

- The value \(0.5(n + 1)\) indicates the position in the ordered data set.
- If \(n\) is even, we choose a value halfway between the two middle observations.

Example: Find the median for the following two sets of measurements

\[
\begin{align*}
2, 9, 11, 5, 6 & \quad 2, 9, 11, 5, 6, 27
\end{align*}
\]
Measures of central tendency for ungrouped data

Mode
- The mode is the value that occurs with the highest frequency in a data set.
- A data with each value occurring only once has no mode.
- A data set with only one value occurring with highest frequency has only one mode, it is said to be unimodal.
- A data set with two values that occurs with the same (highest) frequency has two modes, it is said to be bimodal.
- If more than two values in a data set occur with the same (highest) frequency, it is said to be multimodal.

Example: You are given 8 measurements: 3, 5, 4, 6, 12, 5, 6, 7. Find
   a) The mean.  b) The median.  c) The mode.

Relationships among the mean, median, and Mode
- Symmetric histograms when \[ \text{Mean} = \text{Median} = \text{Mode} \]
- Right skewed histograms when \[ \text{Mean} > \text{Median} > \text{Mode} \]
- Left skewed histograms when \[ \text{Mean} < \text{Median} < \text{Mode} \]
Measures of dispersion for ungrouped data

**Range**
- Data sets may have same center but look different because of the way the numbers are spread out from center.

*Example:*
Company 1: 47 38 35 40 36 45 39
Company 2: 70 33 18 52 27
- Measure of variability can help us to create a mental picture of the spread of the data.
- The range for ungrouped data

\[
\text{Range} = \text{Largest value} - \text{Smallest value}
\]
- The range, like the mean, is highly influenced by outliers.
- The range is based on two values only.

---

**Variance and standard deviation**
- The standard deviation is the most used measure of dispersion. It tells us how closely the values of a data set are clustered around the mean.
- In general, larger values of standard deviation indicate that values of that data set are spread over a relatively larger range around the mean and vice versa.

\[
\text{Population : } \sigma^2 = \frac{\sum x^2 - \left(\frac{\sum x}{N}\right)^2}{N}, \text{ and } \text{sample : } s^2 = \frac{\sum x^2 - \left(\frac{\sum x}{n}\right)^2}{n-1}
\]

\[
\sigma = \sqrt{\sigma^2} \quad \text{and} \quad s = \sqrt{s^2}
\]
- Standard deviation is always non-negative
Measures of dispersion for ungrouped data

Example: Find the standard deviation of the data set in table.

<table>
<thead>
<tr>
<th>MLB team</th>
<th>2002 payroll (millions of dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaheim Angels</td>
<td>62</td>
</tr>
<tr>
<td>Atlanta Braves</td>
<td>93</td>
</tr>
<tr>
<td>New York Yankees</td>
<td>126</td>
</tr>
<tr>
<td>St. Louis Cardinals</td>
<td>75</td>
</tr>
<tr>
<td>Tampa Bay Devil Rays</td>
<td>34</td>
</tr>
</tbody>
</table>

In some situations we may be interested in a descriptive statistics that indicates how large is the standard deviation compared to the mean.

It is very useful when comparing two different samples with different means and standard deviations.

It is given by:

\[ CV = \left( \frac{\sigma}{\mu} \times 100 \right)\% \]
Mean, variance, and standard deviation for grouped data

Mean for grouped data

- Once we group the data, we no longer know the values of individual observations.
- Thus, we find an approximation for the sum of these values.

Mean for population: \( \mu = \frac{\sum mf}{N} \)

Mean for sample: \( \bar{x} = \frac{\sum mf}{n} \)

Where \( m \) is the midpoint and \( f \) is the frequency of a class.

Variance and standard deviation for grouped data

Population: \( \sigma^2 = \frac{\sum m^2 f - (\sum mf)^2}{N} \)

Sample: \( s^2 = \frac{\sum m^2 f - (\sum mf)^2}{n - 1} \)

Where \( m \) is the midpoint and \( f \) is the frequency of a class.

\( \sigma = \sqrt{\sigma^2} \) and \( s = \sqrt{s^2} \)
Example: The table below gives the frequency distribution of the daily commuting times (in minutes) from home to CBA for all 25 students in QMIS 120. Calculate the mean and the standard deviation of the daily commuting times.

<table>
<thead>
<tr>
<th>Daily commuting time (min)</th>
<th>( f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to less than 10</td>
<td>4</td>
</tr>
<tr>
<td>10 to less than 20</td>
<td>9</td>
</tr>
<tr>
<td>20 to less than 30</td>
<td>6</td>
</tr>
<tr>
<td>30 to less than 40</td>
<td>4</td>
</tr>
<tr>
<td>40 to less than 50</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>25</strong></td>
</tr>
</tbody>
</table>

Use of standard deviation

**Z-Scores**

We often are interested in the relative location of a data point \( x_i \) with respect to the mean (how far or how close).

The z-scores (standardized value) can be used to find the relative location of a data point \( x_i \) compared to the center of the data.

It is given by

\[
Z_i = \frac{x_i - \bar{x}}{s}
\]

The z-score can be interpreted as the number of standard deviations \( x_i \) is from the mean.
Use of standard deviation

Example: Find the z-scores for the following data where the mean is 44 and the standard deviation is 8.

<table>
<thead>
<tr>
<th>Number of students in a class</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
</tr>
<tr>
<td>54</td>
</tr>
<tr>
<td>42</td>
</tr>
<tr>
<td>46</td>
</tr>
<tr>
<td>32</td>
</tr>
</tbody>
</table>

Use of standard deviation

Chebyshev’s Theorem

- Chebyshev’s theorem allows you to understand how the value of a standard deviation can be applied to any data set.
- Chebyshev’s theorem: The fraction of any data set lying within $k$ standard deviations of the mean is at least

$$1 - \frac{1}{k^2}$$

where $k$ = a number greater than 1.
- This theorem applies to all data sets, which include a sample or a population.
- Chebyshev’s theorem is very conservative but it can be applied to a data set with any shape.
Use of standard deviation

**Example:** The average systolic blood pressure for 4000 women who were screened for high blood pressure was found to be 187 with standard deviation of 22. Using Chebyshev’s theorem, find at least what percentage of women in this group have a systolic blood pressure between 143 and 231.
Use of standard deviation

**Empirical Rule**

- The empirical rule gives more precise information about a data set than the Chebyshev’s Theorem, however it only applies to a data set that is bell-shaped.

- **Theorem:**
  1- 68% of the observations lie within one standard deviation of the mean.
  2- 95% of the observations lie within two standard deviations of the mean.
  3- 99.7% of the observations lie within three standard deviations of the mean.

---

**Use of standard deviation**

Example: The age distribution of a sample of 5000 persons is bell-shaped with a mean of 40 years and a standard deviation of 12 years. Determine the approximate percentage of people who are 16 to 64 years old.
Use of standard deviation

25

Detecting outliers

- Sometimes a data set may have one or more observation that is unusually small or large value.
- This extreme value is called an outlier and can be detected using the z-score and the empirical rule for data with bell-shape distribution.
- An experienced statistician may face the following situations and need to take an action

<table>
<thead>
<tr>
<th>Outlier</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>A data value that was incorrectly recorded</td>
<td>Correct it before any further analysis</td>
</tr>
<tr>
<td>A data value that was incorrectly included</td>
<td>Remove it before any further analysis</td>
</tr>
<tr>
<td>A data value that belongs to the data set and correctly recorded</td>
<td>Keep it!</td>
</tr>
</tbody>
</table>

QM-120, M. Zainal

Measure of position

26

Quartiles and interquartile range

- A measure of position which determines the rank of a single value in relation to other values in a sample or population.
- Quartiles are three measures that divide a ranked data set into four equal parts

![Diagram showing quartiles](QM-120, M. Zainal)
Measure of position

Calculating the quartiles

- The second quartile is the median of a data set.
- The first quartile, Q₁, is the value of the middle term among observations that are less than the median.
- The third quartile, Q₃, is the value of the middle term among observations that are greater than the median.

Interquartile range (IQR) is the difference between the third quartile and the first quartile for a data set.

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1$$

Example: For the following data

75.3  82.2  85.8  88.7  94.1  102.1  79.0  97.1  104.2  119.3  81.3  77.1

Find

- The values of the three quartiles.
- The interquartile range.
- Where does the 104.2 fall in relation to these quartiles.
Measure of position

Percentile and percentile rank

- Percentiles are the summary measures that divide a ranked data set into 100 equal parts.
- The $k^{th}$ percentile is denoted by $P_k$, where $k$ is an integer in the range 1 to 99.

The approximate value of the $k^{th}$ percentile is

$$P_k = \text{Value of the } \left( \frac{kn}{100} \right)^{th} \text{ term in a ranked data set}$$

Example: For the following data

75.3  82.2  85.8  88.7  94.1  102.1  79.0  97.1  104.2  119.3  81.3  77.1

Find the value of the 40th percentile.

Solution:
Box-and-whisker plot

- Box and whisker plot gives a graphic presentation of data using five measures:
  - $Q_1$, $Q_2$, $Q_3$, smallest, and largest values.
- Can help to visualize the center, the spread, and the skewness of a data set.
- Can help in detecting outliers.
- Very good tool of comparing more than a distribution.
- Detecting an outlier:
  - Lower fence: $Q_1 - 1.5 \times (IQR)$
  - Upper fence: $Q_3 + 1.5 \times (IQR)$
  - If a data point is larger than the upper fence or smaller than the lower fence, it is considered to be an outlier.

To construct a box plot

- Draw a horizontal line representing the scale of the measurements.
- Calculate the median, the upper and lower quartiles, and the IQR for the data set and mark them on the line.
- Form a box just above the line with the right and left ends at $Q_1$ and $Q_3$.
- Draw a vertical line through the box at the location of the median.
- Mark any outliers with an asterisk (*) on the graph.
- Extend horizontal lines called “Whiskers” from the ends of the box to the smallest and largest observation that are not outliers.
Box-and-whisker plot

Example: Construct a box plot for the following data.

340  300  470  340  320  290  260  330

Measures of association between two variables

- So far we have studied numerical methods to describe data with one variable.
- Often decision makers are interested in the relationship between two variables.
- To do so, we will use descriptive measure that is called covariance.
- Covariance assigns a numerical value to the linear relationship between two variables (see scatter diagram). It is given by

\[
\text{Sample covariance: } S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}
\]

\[
\text{Population covariance: } \sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}
\]
Measures of association between two variables

- A big disadvantage of the covariance is that it depends on the units of measurement for x and y.

- For the same data set, we will have two different covariance values depending on the units (i.e. height in meters or centimeters will make a big difference).

- *Pearson’s correlation coefficient* is a good remedy to that problem as it can go only from -1 to 1.

- It is given by

\[
\text{Population correlation coefficient: } \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}
\]

\[
\text{Sample correlation coefficient: } r_{xy} = \frac{S_{xy}}{S_x S_y}
\]

Example: A golfer is interested in investigating the relationship, if any, between driving distance and 18-hole score.

<table>
<thead>
<tr>
<th>Average Driving Distance (meters)</th>
<th>Average 18-Hole Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>277.6</td>
<td>69</td>
</tr>
<tr>
<td>259.5</td>
<td>71</td>
</tr>
<tr>
<td>269.1</td>
<td>70</td>
</tr>
<tr>
<td>267.0</td>
<td>70</td>
</tr>
<tr>
<td>255.6</td>
<td>71</td>
</tr>
<tr>
<td>272.9</td>
<td>69</td>
</tr>
</tbody>
</table>
Example: Find the covariance and Pearson’s correlation coefficient for the following data

<table>
<thead>
<tr>
<th>Week</th>
<th>Number of commercials (x)</th>
<th>Sales in $ (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>57</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>54</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>54</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>63</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>48</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>59</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>46</td>
</tr>
</tbody>
</table>

Measures of association between two variables