Chapter 2 Methods for Describing Sets of Data

2.1 Describing qualitative data

Recall qualitative data: non-numerical or categorical data

**Basic definitions:**

A ____________ is one of the categories into which qualitative data can be classified.

The ______________ is the number of observations in the data set falling into a particular class.

The ________________ is the class frequency divided by the total number of observations in the data set. (in decimal)

Relative Frequency = \( \frac{\text{Frequency}}{n} \)

The ________________ is the class relative frequency multiplied by 100.

Class percentage = (class relative frequency) * 100

**Graphical Descriptive Methods for Qualitative Data:**

__________________________: consists of two columns, one is the classes, the other one is the class frequency.

__________________________: consists of two columns, one is the classes, the other one is the class relative frequency or the class percentage.

__________________________: The categories (classes) of the qualitative variable are represented by bars, where the height of each bar is either __________ or class relative frequency (or class percentage).

__________________________: The categories (classes) of the qualitative variable are represented by slices of a pie (circle). The size of each slice is proportional to the ________________.

__________________________: A bar graph with the categories (classes) of the qualitative variable arranged by height in ____________ order from left to right.
Example 1: Twenty-five army soldiers were given a blood test to determine their blood type.


Construct a frequency and relative frequency distribution table for the data.

Example 2: Road Rage: The following table provides the days on which 69 road rage incidents occurred. Use Descriptive Methods to describe this qualitative data.

1. Summary table: Frequency and relative frequency distribution table.

The following table is from SPSS output.

<table>
<thead>
<tr>
<th>Day of the week</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>18</td>
<td>26.1</td>
<td>26.1</td>
<td>26.1</td>
</tr>
<tr>
<td>M</td>
<td>5</td>
<td>7.2</td>
<td>7.2</td>
<td>33.3</td>
</tr>
<tr>
<td>Sa</td>
<td>7</td>
<td>10.1</td>
<td>10.1</td>
<td>43.5</td>
</tr>
<tr>
<td>Su</td>
<td>5</td>
<td>7.2</td>
<td>7.2</td>
<td>50.7</td>
</tr>
<tr>
<td>Th</td>
<td>11</td>
<td>15.9</td>
<td>15.9</td>
<td>66.7</td>
</tr>
<tr>
<td>Tu</td>
<td>11</td>
<td>15.9</td>
<td>15.9</td>
<td>82.6</td>
</tr>
<tr>
<td>W</td>
<td>12</td>
<td>17.4</td>
<td>17.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>
Questions:
1. Which class has the highest relative frequency? (_______________)
2. What is the percentage that road rage incidents occurred on Friday or Saturday?

2. Bar graph

3. Pie chart

4. Pareto diagram
**Example 3. Age and Gender:** The following bivariate data on age (in years) and gender were obtained from the 50 students in a freshman calculus course.

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Age</th>
<th>Gender</th>
<th>Age</th>
<th>Gender</th>
<th>Age</th>
<th>Gender</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>M</td>
<td>29</td>
<td>F</td>
<td>22</td>
<td>M</td>
<td>23</td>
<td>F</td>
<td>21</td>
<td>F</td>
</tr>
<tr>
<td>20</td>
<td>M</td>
<td>20</td>
<td>M</td>
<td>23</td>
<td>M</td>
<td>44</td>
<td>M</td>
<td>28</td>
<td>F</td>
</tr>
<tr>
<td>42</td>
<td>F</td>
<td>18</td>
<td>F</td>
<td>19</td>
<td>F</td>
<td>19</td>
<td>M</td>
<td>21</td>
<td>F</td>
</tr>
<tr>
<td>21</td>
<td>M</td>
<td>21</td>
<td>M</td>
<td>21</td>
<td>M</td>
<td>21</td>
<td>F</td>
<td>21</td>
<td>F</td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>26</td>
<td>M</td>
<td>21</td>
<td>F</td>
<td>19</td>
<td>M</td>
<td>24</td>
<td>F</td>
</tr>
<tr>
<td>21</td>
<td>F</td>
<td>24</td>
<td>F</td>
<td>21</td>
<td>F</td>
<td>25</td>
<td>M</td>
<td>24</td>
<td>F</td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>19</td>
<td>M</td>
<td>20</td>
<td>F</td>
<td>21</td>
<td>M</td>
<td>24</td>
<td>F</td>
</tr>
<tr>
<td>19</td>
<td>M</td>
<td>25</td>
<td>M</td>
<td>20</td>
<td>F</td>
<td>19</td>
<td>M</td>
<td>23</td>
<td>M</td>
</tr>
<tr>
<td>23</td>
<td>M</td>
<td>19</td>
<td>F</td>
<td>20</td>
<td>F</td>
<td>18</td>
<td>F</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>23</td>
<td>M</td>
<td>22</td>
<td>F</td>
<td>18</td>
<td>F</td>
<td>19</td>
<td>M</td>
</tr>
</tbody>
</table>

1. Contingency table (summary frequency table)

<table>
<thead>
<tr>
<th></th>
<th>Under 21</th>
<th>21-25</th>
<th>Over 25</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Contingency table (summary relative frequency table)

<table>
<thead>
<tr>
<th></th>
<th>Under 21</th>
<th>21-25</th>
<th>Over 25</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Interpret the results.

1. How many female students are under 21 years old out of these 50 students?

2. What percentage of female students are under 21 years old?

3. How many female students are out of these 50 students and what is the percentage?

4. How many students are over 25 years old and what is the relative frequency?
2.2 Graphical methods for describing Quantitative data

____________ can be used to describe quantitative data.

Example1. Test score:

<table>
<thead>
<tr>
<th>score</th>
<th>frequency</th>
<th>Rel. freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;=90</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>80-89</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>70-79</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>&lt;=60</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

*Note: class can’t be very small or very large.

Example2. DVD price: Describe the data by a frequency and relative frequency distribution table.

$210, 219, 214, 197, 224, 219, 199, 199, 208, 209, 215, 199, 212, 212, 219, 210

<table>
<thead>
<tr>
<th>Price</th>
<th>frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>195-200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200-205</td>
<td></td>
<td></td>
</tr>
<tr>
<td>205-210</td>
<td></td>
<td></td>
</tr>
<tr>
<td>210-215</td>
<td></td>
<td></td>
</tr>
<tr>
<td>215-220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>220-225</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: the borderline observation will classify into the next-highest interval. For example: $210 is classified into the 210-215 class.

Three graphical methods for describing quantitative data: Dot plots, stem-and-leaf display, and histogram.

____________: the horizontal axis is a scale for the quantitative variable. Each dot represents one observation of the data. (It shows how the data spread.)

Example2. DVD price: Describe the data by a dot plot.

$210, 219, 214, 197, 224, 219, 199, 199, 208, 209, 215, 199, 212, 212, 219, 210

The dot plot shows most of the prices fall between ____ and ____.
Questions:
1. How many observations in this data set?
2. What percentage of DVD players prices fall between $200 and $220?
3. define the variable x: the DVD players price, find
   
   \[ P(x < 200) = \]
   \[ P(x \leq 200) = \]
   \[ P(200 < x < 220) = \]

________________________: (Here each value is represented by a gross measurement called the _____ and a fine measurement called the _______)

1) Define the stem you wish to choose. (define it in such a way that the number of stems lies between ______________.)

2) List all possible digits of the stem in a column from the ________________.

3) List the leaves next to the appropriate stem (generally, only ____ digit is displayed in the leaf.)

4) _______ the leaves.

Example 2. DVD price: Describe the data by a stem-leaf display.

$210, 219, 214, 197, 224, 219, 199, 199, 208, 209, 215, 199, 212, 212, 219, 210$

The following is a stem-leaf plot from MiniTab.

Stem-and-leaf of price  N = 16
Leaf Unit = 1.0
19 7999
20
20 89
21 00224
21 5999
22 4
Questions:
1. How many observations in this data set?
2. What percentage of DVD players prices fall between $200 and $220?
3. define the variable x: the DVD players price, find
   \[ P(x < 200) = \]
   \[ P(x \leq 200) = \]
   \[ P(200 < x < 220) = \]

*Note: 1. Stem-and-leaf display gives the ___________ of data set, and it also gives the actual values of observations;
2. Stem-leaf display is not appropriate for a __________________ data set.

Example: The following data represents the breaking strengths of 20 linen threads: (in ounces)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>32.5</td>
<td>15.2</td>
<td>29.3</td>
<td>24.5</td>
</tr>
<tr>
<td>21.2</td>
<td>20.0</td>
<td>23.9</td>
<td>33.0</td>
</tr>
<tr>
<td>27.3</td>
<td>41.0</td>
<td>36.8</td>
<td>19.2</td>
</tr>
<tr>
<td>20.6</td>
<td>26.9</td>
<td>28.7</td>
<td>34.2</td>
</tr>
<tr>
<td>25.4</td>
<td>34.6</td>
<td>33.2</td>
<td>37.0</td>
</tr>
</tbody>
</table>

Notes:

a) If there are too _____ leaves per stem, you can split the stem up, using ____ lines per stem.

b) If there are too many trailing digits, we can ______ some of these digits to maximize clarity.

________________________: A graph that displays the classes on the horizontal axis and the frequencies of the classes on the vertical axis.
________________________ histogram: A graph that displays the classes on the horizontal axis and the relative frequencies of the classes on the vertical axis.

Example2. DVD price: Describe the data by a histogram.

$210, 219, 214, 197, 224, 219, 199, 199, 208, 209, 215, 199, 212, 212, 219, 210$
1. How many observations are in this data set?

2. How many DVD player prices fall in $195 and $200, and what is the relative frequency?

3. Which interval with the highest relative frequency? How much is it?

4. What is the relative frequency of class interval ($205, $220)? Interpret it.
Note:
1. The total area under the curve is _______.
2. The proportion of the total _______ under the histogram that falls above a particular interval of the horizontal axis is equal to __________________ of measurements falling in the interval.
3. For a very large data set, when the class intervals become small enough, a relative frequency histogram will appear as a ___________________.

2.3 Summation notation

Measurements:

Sum of measurement:

Sum of squares:

Square of sum:

Example: 5, 3, 8, 5, 4,

* Check the column sums in the following table to find the different summations.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x^2$</th>
<th>$x - 5$</th>
<th>$(x - 5)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.4 Numerical measures of central tendency

Numerical descriptive methods measure two important data characteristics: Central tendency and variability.

The ________________ of a data set: the tendency of the data to cluster, or center, about certain numerical values.

The ________________ of a data set: the spread of the data.

Three most numerical measures of central tendency: Mean, Median, Mode

• __________ of a data set is the arithmetic average of the data set. It measures the central tendency based on the values of observations.

For sample data, sample mean:

Find mean of the two sample data sets.
Example 1: 5, 3, 8, 5, 2, 6, 9

Example 2: Math test score: 89, 91, 73, 76, 69, 88, 79, 84, 85, 81

Example 3: Here is the survey result of prices of 10 DVD players. What is the average price of these 10 DVD players?

$210, $219, $214, $197, $224, $219, $209, $215, $212, $219

Find a weighted mean.

You are taking a class in which your grade is determined from five sources: 50% from your test mean, 15% from your midterm, 20% from your final exam, 10% from your computer lab work, and 5% from your homework. Your scores are 86 (test mean), 96 (midterm), 82 (final exam), 98 (computer lab), and 100 (homework). What is the weighted mean of your scores? If the minimum average for an A is 90, did you get an A?
### Data Analysis

#### Source Table

<table>
<thead>
<tr>
<th>Source</th>
<th>Score, x</th>
<th>Weight, w</th>
<th>x*w</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test mean</td>
<td>86</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>midterm</td>
<td>96</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>Final exam</td>
<td>82</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Computer lab</td>
<td>98</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>homework</td>
<td>100</td>
<td>5%</td>
<td></td>
</tr>
</tbody>
</table>

#### Find the mean from a frequency distribution table.

#### Find the mean number of credit cards the students have.

<table>
<thead>
<tr>
<th># of credit cards</th>
<th>students</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

### Population Mean

Note: The sample mean ___ will play an important role in accomplishing our objective of making inference about population based on sample information. We often use the sample mean ___ to estimate the population mean_____.

Note: How accurate using $\bar{x}$ to estimate $\mu$ depends on:

1. the ______ of the sample. The larger the sample, the more accurate the estimate will tend to be.
2. the ______________, or spread of the data. If all other factors remaining constant, the smaller the variability, the more accurate the estimate.

- _________ $m$: the middle number when the measurements are arranged in ascending (or descending) order. Median divides data set into two parts, 50% of the observations below the median and 50% of observations above the median. It is the __________ of observations.

### How to find a sample median $m$:

Arrange the n measurements from __________ to __________.

1. If n is odd, media $m$ is the ________ number.
The position of the median is: (______)

2. If n is even, media \( m \) is the __________________________ numbers.

The positions of the two middle numbers are: __________

**Find the median of the two data sets.**
Example1: 5, 3, 8, 5, 2, 6, 9

Arrange in increasing order: __________________________

The sample size ________(odd),

so the **position** of the median is____________

the 4\(^{th}\) measurement is the median: _______.

Example2: Math test score: 89, 91, 73, 76, 69, 88, 79, 84, 85, 81

Arrange in increasing order: _____________________________________

The sample size ___________(even),

so the **position** of the median is the middle of _______________________,

The 5\(^{th}\) measurement is 81 and the 6\(^{th}\) is 84, so the median is ____________ .

Example3: Here is the survey result of prices of 10 DVD players. What is the median price of these 10 DVD players?

\[
\begin{align*}
$210, & \quad $219, & \quad $214, & \quad $197, & \quad $224, & \quad $219, & \quad $209, & \quad $215, & \quad $212, & \quad $219
\end{align*}
\]

**Note:** In certain situations, when there are some _______observations in a data set, how the extreme observation has effect on the mean and median?

Example1: 5, 3, 8, 5, 2, 6, 9 (mean = 5.43 and median = 5)
Example2: 5, 3, 8, 5, 2, 6, 90 (mean = _____ and median = ____)

Example 3: (math test score) 89, 91, 73, 76, 69, 88, 79, 84, 85, 81 (mean=81.5, median=82.5)
Example 4: (math test score) 7, 91, 73, 76, 69, 88, 79, 84, 85, 81 (mean=_____, median=____)

From above examples, you can find that mean is more _______ to extreme value than median. The _______ may be a better measure of central tendency than the mean when there are some extreme observations in a data set.
The relationship between mean and median:

Detecting Skewness by Comparing the Mean and the Median

If the data set is skewed to the right, then the median is less than the mean.

![Graph showing right-skewed distribution]

If the data set is symmetric, then the mean equals the median.

![Graph showing symmetric distribution]

If the data set is skewed to the left, then the mean is less than the median.

![Graph showing left-skewed distribution]

Example 1: The mean test score for a class is 72 while the median is 80, what type of distribution most likely describes the shape of the test score?

(________________________)

Example 2. The mean time for the patients stay in the hospital is 15 days while the median is 10 days, what type of distribution most likely describes the shape of the days staying in hospital?

(________________________)

● ________: is the measurement that occurs most frequently in the data set.

Find the mode for the following data.
Example 1: 5, 3, 8, 5, 2, 6, 9
The mode is the observation ________.

Example 2: Math test score: 89, 91, 73, 76, 69, 88, 79, 84, 85, 81

Example 3: Here is the survey result of prices of 10 DVD players. What is the mode price of these 10 DVD players?

$210, $219, $214, $197, $224, $219, $209, $215, $212, $219
Example 4. Road Rage:

<table>
<thead>
<tr>
<th>Day of the week</th>
<th>frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>5</td>
<td>0.0725</td>
</tr>
<tr>
<td>Monday</td>
<td>5</td>
<td>0.0725</td>
</tr>
<tr>
<td>Tuesday</td>
<td>11</td>
<td>0.159</td>
</tr>
<tr>
<td>Wednesday</td>
<td>12</td>
<td>0.174</td>
</tr>
<tr>
<td>Thursday</td>
<td>11</td>
<td>0.159</td>
</tr>
<tr>
<td>Friday</td>
<td>18</td>
<td>0.261</td>
</tr>
<tr>
<td>Saturday</td>
<td>7</td>
<td>0.101</td>
</tr>
</tbody>
</table>

___________ is the mode of the data. (since Friday with the highest frequency)

**Note:**
1. the mode is particularly useful for describing ____________ data. The modal category is simply the category (or class) that occurs most often.
2. Mean, median, or mode: The choice of which measure of central tendency to use will depend on the properties of the data set analyzed and on the application. So, it is vital that you understand how the mean, median and mode are computed.

2.5 Numerical measures of variability (range, Standard deviation)

- The __________ is the largest measurement minus the smallest measurement.

\[ Range = \underline{\quad} \]

**Find the range for the following data set:**

Example 1: \(5, 3, 8, 5, 2, 6, 9,\)

Example 2: Math test score: 89, 91, 73, 76, 69, 88, 79, 84, 85, 81

Example 3: Here is the survey result of prices of 10 DVD players. What is the range of these 10 DVD players’ prices?

\[ $210, $219, $214, $197, $224, $219, $209, $215, $212, $219 \]

Note: range is easy to calculate, and easy to interpret; but it is insensitive when the data set is large.

It is only use the information of the smallest and the largest.

- The ______________ for a sample of \(n\) measurements is equal to the sum of the squared distances from the mean divided by \((n – 1)\).
Note: A shortcut formula for calculating $s^2$ is:

- The ____________________________ is the square root of the sample variance.

Example 1: Calculate mean, the sample variance and sample standard deviation for data 1, 2, 3, 4, 5.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>

Example 2: Calculate mean, the sample variance and sample standard deviation for data 2, 3, 3, 3, 4.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>

Example 3: Here is the survey result of prices of 10 DVD players. What is the variance and standard deviation of these 10 DVD players’ prices?

$210, 219, 214, 219, 224, 219, 209, 215, 212, 219$
Note: 1. The standard deviation is expressed in the ______________ of measurement.
2. The ______the standard deviation, the _______________ the data.

• **The symbols for sample statistic and population parameter:**
  - Sample mean ____                   population mean ____
  - Sample variance ____                population variance____
  - Sample standard deviation ____         population standard deviation ____

Note: 1. The ___________________ can be calculated based on sample data, so it is known while the parameter usually is unknown.
2. _______________ can be used to estimate the corresponding ______________________.
3. The ______ the sample size, the better the estimation.

2.6 Interpreting the standard deviation

How does the standard deviation provide a measure of variability of a data set?

Interpreting the standard deviation: **Chebyshev’s rule** and **Empirical rule**

• _______________ applies to ___ data set, regardless of the shape of the frequency distribution of the data. (at least ______of the measurements will fall within $k$ standard deviation of the mean.)
  a. At least ______ of the measurements will fall within ____ standard deviations of the mean (within interval $(\bar{x} - 2s, \bar{x} + 2s)$ for samples; $(\mu - 2\sigma, \mu + 2\sigma)$ for populations).
  b. At least ______ of the measurements will fall within ____ standard deviations of the mean (within interval $(\bar{x} - 3s, \bar{x} + 3s)$ for samples; $(\mu - 3\sigma, \mu + 3\sigma)$ for populations).

• _______________ is a rule of thumb that applies to data sets with frequency distributions that are mound shaped and ________.
  a. Approximately ______ of the measurements will fall within ____ standard deviation of the mean (within interval $(\bar{x} - s, \bar{x} + s)$ for samples; $(\mu - \sigma, \mu + \sigma)$ for populations).
  b. Approximately ______ of the measurements will fall within ___ standard deviations of the mean
(within interval \((\bar{x} - 2s, \bar{x} + 2s)\) for samples; \((\mu - 2\sigma, \mu + 2\sigma)\) for populations).

c. Approximately _____ of the measurements will fall within ____ standard deviations of the mean
(within interval \((\bar{x} - 3s, \bar{x} + 3s)\) for samples; \((\mu - 3\sigma, \mu + 3\sigma)\) for populations).

**Example 1:** A study was designed to investigate the effects of teaching method on a student's achievement in a mathematics course. These students obtained a mean score of 320 with a standard deviation of 50 on a standardized test. Assuming no information concerning the shape of the distribution is known, what percentage of the students scored between 220 and 420?

**Example 2:** A study was designed to investigate the effects of teaching method on a student's achievement in a mathematics course. These students obtained a mean score of 320 with a standard deviation of 50 on a standardized test. Assuming a mound-shaped and symmetric distribution of the score, what percentage of the students scored between 170 and 470?

**Example 3:** A manufacturer of automobile batteries claims that the average length of life time for its battery is 60 months and the standard deviation is 10 months, what is the minimum percentage of this brand battery that will last within a time interval of 30 months to 90 months?

**Example 4:** A manufacturer of automobile batteries claims that the average length of life time for its battery is 60 months and the standard deviation is 10 months. If it turns out that the distribution of life times for this brand battery is normally distributed (or bell shaped),

1. What percentage of the batteries will last more than 50 months?

2. What percentage of the batteries will last less than 50 months?

3. What is the percentage of the battery that life time will last between 40 and 80 months?

4. What percentage of the batteries will last more than 50 months?
5. What percentage of the batteries will last more than 70 months?

6. What percentage of the batteries will last less than 40 months?

7. Suppose you buy one this brand battery. It lasts less than 40 months. What could you infer about the manufacturer’s claim?

Note:
1. Both rules apply to either _____ data sets or _____________ data set.

2. From these two rules, _____ (at least 75% or approx. 95%) of the measurements will within ___ standard deviations of mean, and ________ (at least 89% or approx. 99.7%) of the measurements will fall within ____ standard deviations of the mean.

2.7 Numerical Measures of relative standing (percentile and z-score)

_____________: descriptive measures of the relationship of a measurement to the rest of the data.

Percentile and z-score are two used to measure of relative standing.

- For any set of n measurements (arranged in _______ order), the ________________ is a number such that p% of the measurements fall below that number and (100-p)% fall above it.

Q1: Ana’s reading score is ranked as 99th percentile in her school. What percentage of students has higher score than Ana?

Q2: This infant’s weight is ranked as 60th percentile. What percentage of infants is lighter than this infant?

Figure 2.23
Location of 90th percentile for yearly sales of oil companies

Figure 2.27
The quartiles for a data set
Three important Percentiles are:

___ percentile (_____________): the median of the data set that lies at or below the median of the entire data set
___ percentile = 2nd Quartile = median
___ percentile (_____________): median of the data set that lies at or above the median of the entire data set.

Example: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Median = 50th percentile = ____

_________ = 25th percentile = ____

_________ = 75th percentile = ____

- A ___________ represents the distance between a given measurement x and the mean, expressed in standard deviations.

Sample z-score for a measurement x is:

Population z-score for a measurement x is:

**Example1:** Find the z-score for the value 96, when the mean is 93 and the standard deviation is 3. (interpret it)

**Example2:** A manufacturer of automobile batteries claims that the average length of life time for its battery is 60 months and the standard deviation is 10 months. Calculate the z-scores for one this kind battery lasts 40 months and another one lasts 90 months.

Interpret: 40 months is____ standard deviations ______ mean 60 months.

Interpret: 90 months is____ standard deviations ________ mean 60 months.
Example 3: A student took three Statistics exams last semester.

<table>
<thead>
<tr>
<th>Exam</th>
<th>Score</th>
<th>Mean of class</th>
<th>Standard deviation of class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam1</td>
<td>76</td>
<td>70</td>
<td>10</td>
</tr>
<tr>
<td>Exam2</td>
<td>83</td>
<td>75</td>
<td>9</td>
</tr>
<tr>
<td>Exam3</td>
<td>76</td>
<td>66</td>
<td>8</td>
</tr>
</tbody>
</table>

Compare to the whole class, which exam did he do best? (compare z-score)

Note: 1. if $z > 0$, it means the observation is ___________ the mean.

2. if $z < 0$, it means the observation is ___________ the mean.

3. if $z = 0$, it means the observation ___________ the mean.

Interpretation of z-score for bell-shaped distributions of data:

Interpretation of z-score for bell-shaped distributions of data:

Note that this interpretation of z-scores is identical to that given by the empirical rule for mound-shaped distributions (Table 2.7). The statement that a measurement falls into the interval from $(\mu - \sigma)$ to $(\mu + \sigma)$ is equivalent to the statement that a measurement has a population z-score between -1 and 1, since all measurements between $(\mu - \sigma)$ and $(\mu + \sigma)$ are within one standard deviation of $\mu$. These z-scores are displayed in Figure 2.28.

1. Approximately _________ of the measurements will have a z-score between -1 and 1.

2. Approximately _________ of the measurements will have a z-score between -2 and 2.

3. Approx. _______ (almost all) of the measurements will have a z-score between -3 and 3.

4. If _________, it means the observation is highly suspected to be an outlier (rare observations).
Understand z-score:
Q1: Ana’s reading score is 1.57 standard deviations above the school average.
Q2: This infant’s weight is 3 standard deviations above the average.
Q3: This baby’s height is 0.16 standard deviation below the average.

Q: The following table is the descriptive statistics for a class test score. (two intervals?)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>80.32203</td>
</tr>
<tr>
<td>Median</td>
<td>84</td>
</tr>
<tr>
<td>Mode</td>
<td>92</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>10.84442</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>117.6014</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>48</td>
</tr>
<tr>
<td>Minimum</td>
<td>52</td>
</tr>
<tr>
<td>Maximum</td>
<td>100</td>
</tr>
<tr>
<td>Sum</td>
<td>4916</td>
</tr>
<tr>
<td>Count</td>
<td>59</td>
</tr>
</tbody>
</table>

Q: Calculate the z-scores for two students: one scored 98, the other one scored 62?

Learning Objective of Chapter 2:
1. Graphical methods to describe qualitative data: frequency and relative frequency distribution table, bar graph, pie chart and pareto diagram. (construct and interpret)
2. Graphical methods to describe quantitative data: frequency and relative frequency distribution table, dot plot, stem-leaf display and histogram. (construct and interpret)
3. Numerical measures to describe the central tendency of quantitative data: mean, median and mode (calculate and interpret; relationship between mean and median (skewness); if there is extreme value, median is less sensitive than mean; mode can describe the qualitative data.)
4. Numerical measures to describe the variability of quantitative data: range, sample variance and sample standard deviation (calculate and interpret, two rules—application restriction, percentage and corresponding intervals)
5. Numerical measures to describe the relative standing of quantitative data: percentile and z-score (percentile---understand and interpret; z-score-- calculate and interpret)