Chapter 2
Descriptive Statistics I:
Tabular and Graphical Presentations

Learning objectives

1. Single variable
   1.1. How to use Tables and Graphs to summarize data
      1.1.1. Qualitative data
      1.1.2. Quantitative data
   1.2. How to use Stem-and-Leaf display to explore data

2. Two variables
   2.1. How to identify and understand potential relationship between variables
      2.1.1. Using Crosstabulation
         – Power of Simpson’s paradox when present
      2.1.2. Using Scatter Diagram and Trendline
L.O. 1.1.1. Table and Graph for Qualitative Data

- Tabular methods
  - Frequency Distribution
  - Relative Frequency Distribution
  - Percent Frequency Distribution

- Graphical methods
  - Bar Graph
  - Pie Chart
A frequency distribution is a tabular summary of data showing the frequency (or number) of items in each of several nonoverlapping classes.

The objective is to provide insights about the data that cannot be quickly obtained by looking only at the original data.

---

Guests staying at Marada Inn were asked to rate the quality of their accommodations as being **excellent**, **above average**, **average**, **below average**, or **poor**. The ratings provided by a sample of 20 guests are:

<table>
<thead>
<tr>
<th>Rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Above Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Above Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Above Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Above Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Above Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

L.O. 1.1.1.
- Frequency
- Bar graph
- Pie chart
### Frequency Distribution

<table>
<thead>
<tr>
<th>Rating</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Percent Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>2</td>
<td>.10</td>
<td>10</td>
</tr>
<tr>
<td>Below Average</td>
<td>3</td>
<td>.15</td>
<td>15</td>
</tr>
<tr>
<td>Average</td>
<td>5</td>
<td>.25</td>
<td>25</td>
</tr>
<tr>
<td>Above Average</td>
<td>9</td>
<td>.45</td>
<td>45</td>
</tr>
<tr>
<td>Excellent</td>
<td>1</td>
<td>.05</td>
<td>5</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>

\[.10(100) = 10\]

\[1/20 = .05\]

### Bar Graph

- **Marada Inn Quality Ratings**

- Frequency
  - Poor
  - Below Average
  - Average
  - Above Average
  - Excellent

- Rating

---

Slide 7

---

Slide 8
Marada Inn Quality Ratings

- Excellent: 5%
- Poor: 10%
- Below Average: 15%
- Average: 25%
- Above Average: 45%

L.O. 1.1.2. Table and Graph for Quantitative Data

- Tabular methods
  - Frequency Distribution
  - Relative Frequency and Percent Frequency Distributions

- Graphical methods
  - Dot Plot
  - Histogram
  - Cumulative Distributions (called Ogive when shown as a line graph)
The manager of Hudson Auto would like to have a better understanding of the cost of parts used in the engine tune-ups performed in the shop. She examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest dollar, are listed on the next slide.

Example: Hudson Auto Repair

Sample of Parts Cost for 50 Tune-ups

<table>
<thead>
<tr>
<th>91</th>
<th>78</th>
<th>93</th>
<th>57</th>
<th>75</th>
<th>52</th>
<th>99</th>
<th>80</th>
<th>97</th>
<th>62</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>69</td>
<td>72</td>
<td>89</td>
<td>66</td>
<td>75</td>
<td>79</td>
<td>75</td>
<td>72</td>
<td>76</td>
</tr>
<tr>
<td>104</td>
<td>74</td>
<td>62</td>
<td>68</td>
<td>97</td>
<td>105</td>
<td>77</td>
<td>65</td>
<td>80</td>
<td>109</td>
</tr>
<tr>
<td>85</td>
<td>97</td>
<td>88</td>
<td>68</td>
<td>83</td>
<td>68</td>
<td>71</td>
<td>69</td>
<td>67</td>
<td>74</td>
</tr>
<tr>
<td>62</td>
<td>82</td>
<td>98</td>
<td>101</td>
<td>79</td>
<td>105</td>
<td>79</td>
<td>69</td>
<td>62</td>
<td>73</td>
</tr>
</tbody>
</table>
Dot Plot for Tune-up Parts Cost

Frequency Distribution

For Hudson Auto Repair, if we choose six classes:

- Approximate Class Width = (109 - 52)/6 = 9.5 ≈ 10

<table>
<thead>
<tr>
<th>Parts Cost ($)</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Percent Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>50-59</td>
<td>2</td>
<td>.04</td>
<td>4</td>
</tr>
<tr>
<td>60-69</td>
<td>13</td>
<td>.26</td>
<td>26</td>
</tr>
<tr>
<td>70-79</td>
<td>16</td>
<td>.32</td>
<td>32</td>
</tr>
<tr>
<td>80-89</td>
<td>7</td>
<td>.14</td>
<td>14</td>
</tr>
<tr>
<td>90-99</td>
<td>7</td>
<td>.14</td>
<td>14</td>
</tr>
<tr>
<td>100-109</td>
<td>5</td>
<td>.10</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>1.00</td>
<td>100</td>
</tr>
</tbody>
</table>
Histogram

Tune-up Parts Cost

<table>
<thead>
<tr>
<th>Parts Cost ($)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>50-59</td>
<td>50</td>
</tr>
<tr>
<td>60-69</td>
<td>4</td>
</tr>
<tr>
<td>70-79</td>
<td>12</td>
</tr>
<tr>
<td>80-89</td>
<td>4</td>
</tr>
<tr>
<td>90-99</td>
<td>6</td>
</tr>
<tr>
<td>100-110</td>
<td>2</td>
</tr>
</tbody>
</table>

Histogram: Check the Skewness

- L.O. 1.1.2.
  - Dot plot
  - Frequency
  - Histogram
  - Cumulative Distribution

- Skewed left
  - Left tail is longer than the right tail
  - Example: exam scores

- Skewed right
  - Right tail is longer than the left tail
  - Example: heights of people

- Symmetrical
  - No skewness
  - Example: age of people
Discussion item

- Use the handout given and focus on the portion instructed by me.
- Identify and understand:
  - what is being represented by the Histogram?
  - is descriptive statistics helping these authors to make their case? How?

Cumulative Distributions

- Cumulative frequency distribution – shows the number of items with values less than or equal to the upper limit of each class.
- Cumulative relative frequency distribution – shows the proportion of items with values less than or equal to the upper limit of each class.
- Cumulative percent frequency distribution – shows the percentage of items with values less than or equal to the upper limit of each class.
Cumulative Distributions

Hudson Auto Repair

<table>
<thead>
<tr>
<th>Cost ($)</th>
<th>Cumulative Frequency</th>
<th>Cumulative Relative Frequency</th>
<th>Cumulative Percent Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 59</td>
<td>2</td>
<td>.04</td>
<td>4</td>
</tr>
<tr>
<td>&lt; 69</td>
<td>15</td>
<td>.30</td>
<td>30</td>
</tr>
<tr>
<td>&lt; 79</td>
<td>31</td>
<td>2 + 13</td>
<td>.62</td>
</tr>
<tr>
<td>&lt; 89</td>
<td>38</td>
<td>15/50</td>
<td>62</td>
</tr>
<tr>
<td>&lt; 99</td>
<td>45</td>
<td>.76</td>
<td>.90</td>
</tr>
<tr>
<td>&lt; 109</td>
<td>50</td>
<td>90</td>
<td>100</td>
</tr>
</tbody>
</table>

Ogive with Cumulative Percent Frequencies

Tune-up Parts Cost

(89.5, 76)
L.O. 1.2. Exploratory Data Analysis

- The techniques of exploratory data analysis consist of simple arithmetic and easy-to-draw pictures that can be used to summarize data quickly.

- One such technique is the stem-and-leaf display.
  - A stem-and-leaf display shows both the rank order and shape of the distribution of the data.

---

Stem-and-Leaf Display:
Hudson Auto Example

<table>
<thead>
<tr>
<th>5</th>
<th>2, 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2, 2, 2, 5, 6, 7, 8, 8, 8, 9, 9, 9</td>
</tr>
<tr>
<td>7</td>
<td>1, 1, 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 9, 9, 9</td>
</tr>
<tr>
<td>8</td>
<td>0, 0, 2, 3, 5, 8, 9</td>
</tr>
<tr>
<td>9</td>
<td>1, 3, 7, 7, 8, 9</td>
</tr>
<tr>
<td>10</td>
<td>1, 4, 5, 5, 9</td>
</tr>
</tbody>
</table>
Stretched Stem-and-Leaf Display

If we believe the original stem-and-leaf display has condensed the data too much, we can stretch the display by using two stems for each leading digit(s).

Whenever a stem value is stated twice, the first value corresponds to leaf values of 0 – 4, and the second value corresponds to leaf values of 5 – 9.

5  2
5  7
6  2 2 2 2
6  5 6 7 8 8 8 9 9 9
7  1 1 2 2 3 4 4
7  5 5 5 6 7 8 9 9 9
8  0 0 2 3
8  5 8 9
9  1 3
9  7 7 7 8 9
10  1 4
10  5 5 9
Stem-and-Leaf Display

- Leaf Units
  - A single digit is used to define each leaf.
  - In the preceding example, the leaf unit was 1.
  - Leaf units may be 100, 10, 1, 0.1, and so on.
  - Where the leaf unit is not shown, it is assumed to equal 1.

L.O. 1.2. Crosstabulations and Scatter Diagrams

- Thus far we have focused on methods that are used to summarize the data for one variable at a time.
- Often a manager is interested in tabular and graphical methods that will help understand the relationship between two variables.
- Crosstabulation and a scatter diagram are two methods for summarizing the data for two (or more) variables simultaneously.
L.O. 2.1.1. Crosstabulation

- A **crosstabulation** is a tabular summary of data for two variables.
- **Crosstabulation** can be used when:
  - one variable is qualitative and the other is quantitative,
  - both variables are qualitative, or
  - both variables are quantitative.
- The left and top margin labels define the classes for the two variables.

---

### Example: Finger Lakes Homes

The number of Finger Lakes homes sold for each style and price for the past two years is shown below.

<table>
<thead>
<tr>
<th>Price Range</th>
<th>Colonial</th>
<th>Log</th>
<th>Split</th>
<th>A-Frame</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; $99,000</td>
<td>18</td>
<td>6</td>
<td>19</td>
<td>12</td>
<td>55</td>
</tr>
<tr>
<td>&gt; $99,000</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>30</strong></td>
<td><strong>20</strong></td>
<td><strong>35</strong></td>
<td><strong>15</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

---

*Frequency distribution for the price variable*

*Frequency distribution for the home style variable*
Discussion item

- Use the handout given and focus on the portion instructed by me.
- Identify and understand:
  - what is being represented by Table 2?
  - What is being represented by Tables 3 and 4? Are these examples of crosstabulations?

Crosstabulation: Simpson’s Paradox

Aggregate table shows Judge Kendall is better.

<table>
<thead>
<tr>
<th>Verdict</th>
<th>Luckett</th>
<th>Kendall</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upheld</td>
<td>129 (86%)</td>
<td>110 (88%)</td>
<td>239</td>
</tr>
<tr>
<td>Reversed</td>
<td>21 (14%)</td>
<td>15 (12%)</td>
<td>36</td>
</tr>
<tr>
<td>Total (%)</td>
<td>150 (100%)</td>
<td>125 (100%)</td>
<td>275</td>
</tr>
</tbody>
</table>

Judge Luckett

<table>
<thead>
<tr>
<th>Verdict</th>
<th>Common Pleas</th>
<th>Municipal Court</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upheld</td>
<td>29 (91%)</td>
<td>100 (85%)</td>
<td>129</td>
</tr>
<tr>
<td>Reversed</td>
<td>3 (9%)</td>
<td>18 (15%)</td>
<td>21</td>
</tr>
<tr>
<td>Total (%)</td>
<td>32 (100%)</td>
<td>118 (100%)</td>
<td>150</td>
</tr>
</tbody>
</table>

Judge Kendall

<table>
<thead>
<tr>
<th>Verdict</th>
<th>Common Pleas</th>
<th>Municipal Court</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upheld</td>
<td>90 (90%)</td>
<td>20 (80%)</td>
<td>110</td>
</tr>
<tr>
<td>Reversed</td>
<td>10 (10%)</td>
<td>5 (20%)</td>
<td>15</td>
</tr>
<tr>
<td>Total (%)</td>
<td>100 (100%)</td>
<td>25 (100%)</td>
<td>125</td>
</tr>
</tbody>
</table>

However, according to the column percentage tables above, Judge Luckett seems to be better than Judge Kendall.
A scatter diagram is a graphical presentation of the relationship between two quantitative variables. One variable is shown on the horizontal axis and the other variable is shown on the vertical axis. The general pattern of the plotted points suggests the overall relationship between the variables. A trendline is an approximation of the relationship.
Scatter Diagram

- A Negative Relationship

Scatter Diagram

- No Apparent Relationship
Example: Panthers Football Team

- Scatter Diagram
  The Panthers football team is interested in investigating the relationship, if any, between interceptions made and points scored.

<table>
<thead>
<tr>
<th>x = Number of Interceptions</th>
<th>y = Number of Points Scored</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
</tr>
</tbody>
</table>

Scatter Diagram

\[ y \] 
\[ x \] 
\[ 0, 5, 10, 15, 20, 25, 30, 35 \] 
\[ 0, 1, 2, 3, 4 \] 

Number of Interceptions 
Number of Points Scored
Example: Panthers Football Team

- Insights Gained from the Preceding Scatter Diagram
  - The scatter diagram indicates a positive relationship between the number of interceptions and the number of points scored.
  - Higher points scored are associated with a higher number of interceptions.
  - The relationship is not perfect; all plotted points in the scatter diagram are not on a straight line.

Tabular and Graphical Procedures

Data

- Qualitative Data
  - Tabular Methods
    - Frequency Distribution
    - Rel. Freq. Dist.
    - Percent Freq. Distribution
    - Crosstabulation
  - Graphical Methods
    - Bar Graph
    - Pie Chart

- Quantitative Data
  - Tabular Methods
    - Frequency Distribution
    - Rel. Freq. Dist.
    - Cum. Rel. Freq. Distribution
    - Stem-and-Leaf Display
    - Crosstabulation
  - Graphical Methods
    - Dot Plot
    - Histogram
    - Ogive
    - Scatter Diagram
End of Chapter 2